

# Validating RDF with Shape Expressions

Iovka Boneva  
LINKS, INRIA & CNRS  
University of Lille, France

Jose Emilio Labra Gayo  
University of Oviedo  
Spain

Samuel Hym  
2XS  
University of Lille, France

Eric G. Prud'hommeau  
W3C  
Stata Center, MIT

Harold Solbrig  
Mayo Clinic College of Medicine  
Rochester, MN, USA

Slawek Staworko\*  
LINKS, INRIA & CNRS  
University of Lille, France

## Abstract

We propose shape expression schema (ShEx), a novel schema formalism for describing the topology of an RDF graph that uses regular bag expressions (RBEs) to define constraints on the admissible neighborhood for the nodes of a given type. We provide two alternative semantics, multi- and single-type, depending on whether or not a node may have more than one type. We study the expressive power of ShEx and study the complexity of the validation problem. We show that the single-type semantics is strictly more expressive than the multi-type semantics, single-type validation is generally intractable and multi-type validation is feasible for a small class of RBEs. To further curb the high computational complexity of validation, we propose a natural notion of determinism and show that multi-type validation for the class of deterministic schemas using single-occurrence regular bag expressions (SORBEs) is tractable. Finally, we consider the problem of validating only a fragment of a graph with preassigned types for some of its nodes, and argue that for deterministic ShEx using SORBEs, multi-type validation can be performed efficiently and single-type validation can be performed with a single pass over the graph.

## 1 Introduction

Schemas have a number of important functions in databases. They describe the structure of the database, and its knowledge is essential to any user trying to formulate and execute a query or an update over a database instance. Typically schemas allow for effective (and often efficient) algorithms for validating the conformance of a database instance with a given schema. Schemas can also capture the intended meaning of the data stored in database instances and are important for static analysis tasks such as query optimization.

Relational and XML databases have a number of well-established and widely accepted schema formalisms e.g, the SQL Data Definition Language for relational databases and W3C XML Schema or RELAX NG for XML databases. The RDF data model is schema-free in its

---

\*Contact author: [slawomir.staworko@inria.fr](mailto:slawomir.staworko@inria.fr)

conception and in several of its usages. For instance, the linked open data initiative<sup>1</sup> could not have the same success if published data has to comply with a rigid schema. However, more classical applications might need to rely on the fact that the data satisfies some constraints. For instance, one might want to guarantee that an RDF node representing a person has a unique date-of-birth property. Currently, there does not exist any schema formalism for graph databases that addresses that need. RDF Schema [5] is essentially an ontology language and falls short of allowing to describe involved structural properties of the RDF graph. In September 2013, W3C conveyed an industry workshop on RDF validation, the primary outcome of which was an agreement to develop a 'declarative definition of the structure of a graph for validation and description.'<sup>2</sup>

In the present paper, we address this need and propose Shape Expression Schema (ShEx), a formalism under development at W3C as a candidate schema formalism for RDF. A Shape Expression Schema (ShEx) allows to define a set of types that impose structural constraints on nodes and their immediate neighborhood. Figure 2 presents an example of a Shape Expression Schema for RDF database (Figure 1) storing bug reports.

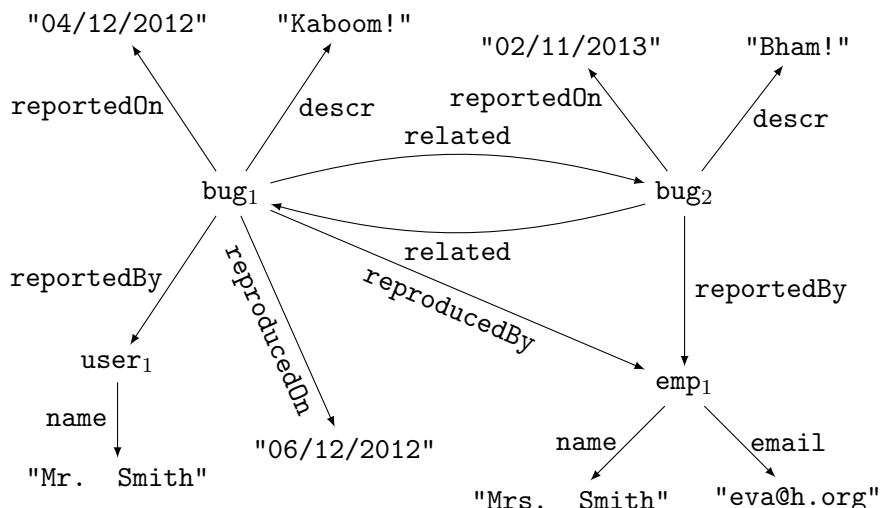


Figure 1: An example of an RDF graph.

Essentially, it says that a bug report has a description, a user who reported it, and on what date. Optionally, a bug report may also have an employee who successfully reproduced the bug, and on what date. Finally, a bug report can have a number of related bug reports. For every user we wish to store his/her name and optionally email. For an employee we wish to store his/her name, either as one string or split into the first and last name, and email address.

In the present paper, we ignore data values and study ShEx from the point of view of the graph topology they are able to define. Note that considering data values would not bring further complexity for the problems we are interested in, namely semantics, expressiveness and complexity of validation of ShEx. A Shape Expression Schema defines a set of types to be associated to graph nodes. Each type defines the admissible collection of outgoing edges and

<sup>1</sup><http://linkeddata.org/>

<sup>2</sup>See <http://www.w3.org/2012/12/rdf-val/report> for a report on that workshop.

```

<BugReport> {
  descr xsd:string,
  reportedBy @<User>,
  reportedOn xsd:dateTime,
  (reproducedBy @<Employee>,
   reproducedOn xsd:dateTime)?
  related @<BugReport>*
}

<User> {
  name xsd:string,
  email xsd:string?
}

<Employee> {
  (name xsd:string | (first-name xsd:string, last-name xsd:string)),
  email xsd:string,
}

```

Figure 2: An example of a Shape Expression Schema.

the types of the nodes they lead to. Naturally, such a schema bears a strong resemblance to RELAX NG and DTDs which also use regular expressions to describe the allowed collections of children of an XML node. The most important difference comes from the fact that in XML, the children of a node are ordered, and the regular expressions in DTDs and RELAX NG schemas define admissible sequences of children, whereas for RDF graphs, no order on the neighborhood of a given node can be assumed. As the regular expressions used in *ShEx* define bags (multisets) of symbols rather than sequences, we call them *regular bag expressions* (RBEs).

The semantics of Shape Expression Schemas is quite natural. An RDF graph is valid if it is possible to assign types to the nodes of the graph in a manner that satisfies all shape expressions of the schema. A natural question arises: can a node be assigned more than one type? In most applications the *multi-type semantics*, which permits assigning multiple types to a node, seems to be more natural. For instance, the RDF graph in Figure 1 requires assigning both the type *User* and the type *Emp* to the node *emp<sub>1</sub>* because *emp<sub>1</sub>* has reported *bug<sub>2</sub>* and reproduced *bug<sub>1</sub>*. However, there are applications where the single-type semantic may be more suitable e.g., when modeling graph transformations that visit every node exactly once.

Since any single-type typing is also a multi-type typing, one could (mistakenly) believe that the single-type semantic is weaker than the multi-type semantics. We show, however, the converse: single-type semantic is in fact strictly more expressive than the multi-type semantics. We also compare the expressive power of *ShEx* with two standard yardstick logics for graphs: first-order logic (FO) and existential monadic second-order logic ( $\exists$ MSO). *ShEx* are not comparable with FO. Because *ShEx* are able to define arbitrary Presburger constraints, they are not comparable with MSO. However,  $\exists$ MSO captures *ShEx* that use restricted regular

expressions. Then, we compare the expressive power of **ShEx** with graph grammars and languages of graphs definable with homomorphisms. We also show that **ShEx** under both semantics are closed under union and intersection but not under complement.

Next, we study the complexity of the validation problem i.e., checking if a given graph has a valid typing w.r.t. the given schema. Naturally, this problem comes in two flavors, depending on the chosen semantic, and we show significant computational differences between them. While validation for both semantics is generally intractable, the multi-type semantics admits a simple and possibly practical class of RBEs with tractable validation. More interestingly, however, we show that the complexity of multi-type validation for **ShEx** using a class  $\mathcal{C}$  of RBEs is closely related to the complexity of the satisfiability problem for  $\mathcal{C}$  with intersection. We study this problem and show that in general this problem is NP-complete. This stands in contrast with the analogue for regular word expressions: it is known that the problem of emptiness of regular expressions with intersection is PSPACE-complete [?].

To lower the complexity of validation, we introduce the notion of determinism. Essentially, determinism requires that every shape expression uses exactly one type with every label. The shape expressions in Figure 2 are deterministic but the following shape expression is not.

```
<BugReport> {
  descr xsd:string,
  (reportedBy @<User> | reportedBy @<Employee>),
  reportedOn xsd:dateTime,
  (reproducedBy @<Employee>,
   reproducedOn xsd:dateTime)?
  related @<BugReport>*
}
```

This shape expression is not deterministic because **reportedBy** is used with two types: **User** and **Employee**. For deterministic shape expression schemas, we are able to relate the complexity of multi-type validation to the problem of checking membership of a bag of symbols to the language of RBEs. While this problem is known to be NP-complete [11], it is generally simpler than the satisfiability problem, and a number of tractable subclasses has already been identified [2, 11]. All known tractable classes of RBEs require the expressions to be single-occurrence i.e., every symbol of the alphabet is used at most once in an RBE. In the present paper, we show that the full class of *single-occurrence regular bag expressions* (SORBE) has in fact tractable membership. Finally, we consider the problem of validating only a fragment of a graph with preassigned types for its root nodes and argue that for deterministic **ShEx** using SORBEs, multi-type validation can be performed efficiently, and show that single-type validation can be performed with a single pass over the graph.

The main contributions of the present paper are:

1. We propose a novel schema formalism for RDF databases and propose two alternative semantics, multi-type and single-type, depending on whether or not a node may have more than one type.
2. We present a thorough analysis of the expressive power of **ShEx** and study their basic properties. We show that the single-type semantic is more expressive than the multi-type semantics.

3. We provide a comprehensive understanding of the complexity of validation and show a very close relationship to the complexity of the problem of satisfiability of regular bag expressions with intersection. We show that satisfiability of arbitrary RBEs with intersection is NP-complete and identify a tractable subclass.
4. We propose a notion of determinism for ShEx that allows to curb the complexity of validation and essentially (Turing) reduces validation to checking membership of a bag to the language of RBE. Additionally, we show that single-occurrence regular bag expressions (SORBE) enjoy a tractable membership problem which makes them a perfect candidate for a practical fragment of deterministic shape expression schemas.

**Organization.** In Section 2 we present basic notions. In Section 3 we introduce Shape Expression Schemas (ShEx) and define the single- and multi-type semantics and in Section 4 we study their closure under Boolean operations. In Section 5 we analyze the expressive power of ShEx. In Section 6 we study the complexity of the validation problem for ShEx. In Section 7 we investigate the problem of satisfiability of regular bag expressions with intersection and characterize its complexity. In Section 8 we introduce a natural notion of determinism for ShEx and identify a rich class of single-occurrence RBEs that together render multi-type validation tractable. In Section 9 we present a preliminary series of experiments performed to validate our algorithms. While in the main part of the paper we consider graphs without data value and with edges labeled by elements from a finite set, in Section 10 we show how to use the presented developments and extend our approach to a graph model that corresponds to RDF more closely and in particular considers data values (literals) and edges consisting of triples of objects. We discuss related work in Section 11. Finally, we conclude and discuss related and future work in Section 12.

## 2 Preliminaries

### 2.1 Regular bag expressions

A *regular bag expression* (RBE) defines bags by using disjunction “|”, unordered concatenation “||”, and unordered Kleene star “\*”. Formally, RBEs over  $\Delta$  are defined with the following grammar:

$$E ::= \epsilon \mid a \mid E^* \mid (E \mid E) \mid (E \parallel E),$$

where  $a \in \Delta$ . The semantics of RBEs is defined as follows:

$$\begin{aligned} L(\epsilon) &= \{\epsilon\}, & L(E_1 \mid E_2) &= L(E_1) \cup L(E_2), \\ L(a) &= \{\{a\}\}, & L(E_1 \parallel E_2) &= L(E_1) \uplus L(E_2), \\ L(E^*) &= \bigcup_{i \geq 0} L(E)^i. \end{aligned}$$

We use two standard macros:  $E^? := (\epsilon \mid E)$  and  $E^+ := (E \parallel E^*)$ . We also use *intervals* on symbols  $a^{[n;m]}$ , where  $n \in \mathbb{N}$  and  $m \in \mathbb{N} \cup \{\infty\}$ , with the natural semantic:  $L(a^{[n;m]}) = \bigcup_{n \leq i \leq m} L(a)^i$ . In general, an *interval*  $[n;m]$  is a finite representation of the set  $\{i \mid n \leq i \leq m\}$ . In this view, an interval  $[n;m]$  is empty if  $m < n$  and in the sequel, we use  $\emptyset$  to denote (the equivalence class of) all empty intervals. Also, the intersection of two intervals can be obtained easily  $[n_1;m_1] \cap [n_2;m_2] = [\max\{n_1, n_2\}; \min\{m_1, m_2\}]$  and the *component-wise addition*  $A \oplus B = \{a + b \mid a \in A, b \in B\}$  can be implemented on intervals as

$[n_1; m_1] \oplus [n_2; m_2] = [n_1 + n_2; m_1 + m_2]$  with  $m + \infty = \infty + m = \infty$  for any  $m \in \mathbb{N} \cup \{\infty\}$ ; also, note that  $\emptyset \oplus [n; m] = [n; m] \oplus \emptyset = \emptyset \oplus \emptyset = \emptyset$ .

We use different syntactic restrictions on RBEs, by indicating which of the syntactic ingredients can be used, among the following:  $a^M$  means that multiplicities among  $\{1, ?, *, +\}$  can be used only on symbols;  $a^I$  means that arbitrary intervals on symbols can be used;  $\parallel$ ,  $|$ , and  $*$  mean that the corresponding operator can be used. For instance,  $\text{RBE}(a^M, \parallel, |)$  is the family of bag languages defined by RBEs that allow multiplicities only on symbols, and unrestricted use of the operators  $\parallel$  and  $|$  e.g.,  $(a^? | b) \parallel c^*$ ;  $\text{RBE}(a, \parallel, |, *)$  is the class of all RBE. Finally, by  $\text{RBE}_1$  we denote the RBEs of the form (with  $a_{i,j} \in \Delta$ )

$$(a_{1,1} | \dots | a_{1,k_1}) \parallel \dots \parallel (a_{n,1} | \dots | a_{n,k_n}).$$

In the sequel, we also use RBE to denote the family of bag languages definable with regular bag expressions, and it should be clear from the context whether “RBE” stands for the class of expressions, or for the class of languages. A number of important facts are known about RBE: it is closed under intersection, union, and complement [14], testing membership  $w \in L(E)$  is NP-complete [11], and so is testing the emptiness of  $L(E_1) \cap L(E_2)$  [6]. It is also known that when a bag of symbols is viewed as vector of natural numbers (obtained by fixing some total order on  $\Delta$ ), the class RBE is equivalent to the class of semilinear sets and the class of vectors definable with Presburger arithmetic [8, 16].

## 2.2 Graphs

We use edge labeled graphs to model RDF databases. We assume a finite set  $\Sigma$  that we use to label the edges of the graphs. An *edge-labeled graph* (or simply a *graph*) is a pair  $G = (V, E)$ , where  $V$  is a finite set of nodes and  $E \subseteq V \times \Sigma \times V$  is the set of edges. The *bag of outbound labels* of  $n$  in  $G$  is the bag

$$\text{out-lab}_G(n) = \{a \mid (n, a, m) \in E\},$$

or more precisely,  $[\text{out-lab}_G(n)](a) = |\{m \in V \mid (n, a, m) \in E\}|$ . The *labeled outbound neighbourhood* of  $n$  in  $G$  is the set

$$\text{out-lab-node}_G(n) = \{(a, m) \in \Sigma \times V \mid (n, a, m) \in E\}.$$

A number of examples of graphs are given in Figure 3.

## 3 Shape Expression Schemas

In this section we formally introduce shape expressions schemas and propose two semantics that we study in the remainder of the paper. We assume a finite set of edge labels  $\Sigma$  and a finite set of types  $\Gamma$ . A *shape expression* is an RBE over  $\Sigma \times \Gamma$ . In the sequel we write  $(a, t) \in \Sigma \times \Gamma$  simply as  $a :: t$ . A *shape expression schema* (ShEx), or simply *schema*, is a tuple  $S = (\Sigma, \Gamma, \delta)$ , where  $\Sigma$  is a finite set of edge labels,  $\Gamma$  is a finite set of types, and  $\delta$  is a *type definition* function that maps elements of  $\Gamma$  to bag languages over  $\Sigma \times \Gamma$ . We only use shape expressions for defining bag languages of  $\delta$ . Typically, we present a ShEx as a collection of rules of the form  $t \rightarrow E$  to indicate that  $\delta(t) = L(E)$ , where  $t \in \Gamma$  and  $E$  is a shape expression over  $\Sigma$  and  $\Gamma$  (naturally, no two rules shall have the same left-hand side). If for some type  $t$

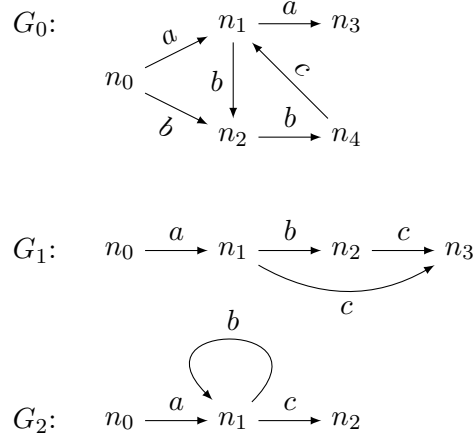


Figure 3: Edge-labeled oriented graphs.

a rule is missing, the default rule is  $t \rightarrow \epsilon$ . For a class of RBEs  $\mathcal{C}$  by  $\text{ShEx}(\mathcal{C})$  we denote the class of shape expression schemas using only shape expressions in  $\mathcal{C}$ . Two example schemas follow:

$$\begin{array}{ll}
 S_0 : t_0 \rightarrow a :: t_1 \parallel b :: t_2 & S_1 : t_0 \rightarrow a :: t_1 \\
 t_1 \rightarrow (a :: t_1 \mid b :: t_2)^* & t_1 \rightarrow b :: t_2 \parallel c :: t_3 \\
 t_2 \rightarrow b :: t_2 \mid c :: t_1 & t_2 \rightarrow (b :: t_2)^? \parallel c :: t_3 \\
 & t_3 \rightarrow \epsilon
 \end{array}$$

The semantics of **ShEx** is natural: graph is valid if it is possible to assign types to the nodes of the graph in a manner that satisfies all shape expressions of the schema. Two variants of semantic can be envisioned depending on whether or not more than one can be assigned to a node.

### 3.1 Single-type semantics

We fix a graph  $G = (V, E)$  and a schema  $S = (\Sigma, \Gamma, \delta)$ . A *single-type typing* (or simply an *s-typing*) of  $G$  w.r.t.  $S$  is a function  $\lambda : V \rightarrow \Gamma$  that associates with every node  $n \in V$  its type  $\lambda(n)$ . An example of an s-typing of  $G_0$  w.r.t.  $S_0$  is

$$\begin{array}{lll}
 \lambda_0(n_0) = t_0, & \lambda_0(n_1) = t_1, & \lambda_0(n_2) = t_2, \\
 \lambda_0(n_3) = t_1, & \lambda_0(n_4) = t_2. &
 \end{array}$$

Next, we identify the conditions that an s-typing needs to satisfy. Given a typing  $\lambda$  and a node  $n \in V$  we define the labeled and typed out-neighborhood of  $n$  w.r.t.  $\lambda$  as bag over  $\Sigma \times \Gamma$ :

$$out\text{-}lab\text{-}type_G^\lambda(n) = \{a :: \lambda(m) \mid (n, a, m) \in E\}.$$

or more precisely,  $[out\text{-}lab\text{-}type_G^\lambda(n)](a :: t) = |\{m \in V \mid (n, a, m) \in E, \lambda(m) = t\}|$ . For instance, for the graph  $G_0$  (Figure 3) and the typing  $\lambda_0$  we have

$$\begin{array}{l}
 out\text{-}lab\text{-}type_{G_0}^{\lambda_0}(n_1) = \{a :: t_1, b :: t_2\}, \\
 out\text{-}lab\text{-}type_{G_0}^{\lambda_0}(n_4) = \{c :: t_1\}.
 \end{array}$$

Now,  $\lambda$  is a *valid* s-typing of  $S$  on  $G$  if and only if every node satisfies the type definition of its associated type i.e., for every  $n \in V$ ,  $out\text{-}lab\text{-}type_G^\lambda(n) \in \delta(\lambda(n))$ . By  $L_s(S)$  we denote the family of all graphs that have a valid s-typing w.r.t. the shape expression schema  $S$ . For a class  $\mathcal{C}$  of bag languages by  $\text{ShEx}_s(\mathcal{C})$  we denote the class of graph languages definable under the single-type semantics with shape expression schemas using shape expressions from  $\mathcal{C}$  only.

Naturally,  $\lambda_0$  is a valid typing of  $G_0$  w.r.t.  $S_0$ .  $G_1$  also has a valid s-typing of  $S_1$ :

$$\lambda_1(n_0) = t_0, \quad \lambda_1(n_1) = t_1, \quad \lambda_1(n_2) = t_2, \quad \lambda_1(n_3) = t_3.$$

$G_2$ , however, does not have a valid s-typing w.r.t.  $S_1$ .

### 3.2 Multi-type semantics

Again, we assume a fixed graph  $G = (V, E)$  and a fixed schema  $S = (\Sigma, \Gamma, \delta)$ . A *multi-type typing* (or simply an *m-typing*) of  $G$  w.r.t.  $S$  is a function  $\lambda : V \rightarrow 2^\Gamma$  that associates with every node of  $G$  a set of types. For instance, an m-typing of  $G_2$  w.r.t.  $S_1$  is

$$\lambda_2(n_0) = \{t_0\}, \quad \lambda_2(n_1) = \{t_1, t_2\}, \quad \lambda_2(n_2) = \{t_3\}.$$

The labeled and typed out-neighborhood of a node is defined in the same way but note that this time it is a bag over  $\Sigma \times 2^\Gamma$ . For instance,

$$out\text{-}lab\text{-}type_{G_2}^{\lambda_2}(n_1) = \{b :: \{t_1, t_2\}, c :: \{t_3\}\}.$$

Now, a *flattening* of a bag over  $\Sigma \times 2^\Gamma$  is any bag over  $\Sigma \times \Gamma$  obtained by choosing one type from every occurrence of every set. For instance,  $out\text{-}lab\text{-}type_{G_2}^{\lambda_2}(n_1)$  has two flattenings:  $\{b :: t_1, c :: t_3\}$  and  $\{b :: t_2, c :: t_3\}$ . Formally, a flattening of a bag  $w$  over  $\Sigma \times 2^\Gamma$  is any bag in the language of the following RBE<sub>1</sub> expression

$$Flatten(w) = \parallel_{a::T \in w} (|_{t \in T} a :: t)$$

where  $a :: T \in w$  indicates that the symbol  $a :: T$  is to be considered  $w(a :: T)$  times. For instance,

$$Flatten(\{a :: \{t_0, t_1\}, a :: \{t_0, t_1\}, b :: t_1\}) = (a :: t_0 \mid a :: t_1) \parallel (a :: t_0 \mid a :: t_1) \parallel (b :: t_1).$$

By  $Out\text{-}lab\text{-}type_G^\lambda(n)$  we denote the set of all flattenings of  $out\text{-}lab\text{-}type_G^\lambda(n)$  i.e.

$$Out\text{-}lab\text{-}type_G^\lambda(n) = L(Flatten(out\text{-}lab\text{-}type_G^\lambda(n))).$$

For instance,

$$Out\text{-}lab\text{-}type_{G_2}^{\lambda_2}(n_1) = \{\{b :: t_1, c :: t_3\}, \{b :: t_2, c :: t_3\}\}.$$

Note that while  $Flatten(out\text{-}lab\text{-}type_G^\lambda(n))$  is an expression of size polynomial in the size of  $G$  and  $S$ , the cardinality of the set  $Out\text{-}lab\text{-}type_G^\lambda(n)$  may be exponential in the size of  $G$  and  $S$ .

Now,  $\lambda$  is a *valid* m-typing of  $G$  w.r.t.  $S$  if and only if:

1. it assigns at least one type to every node,  $\lambda(n) \neq \emptyset$  for  $n \in V$ ,
2. every node satisfies the type definition of every type assigned to the node i.e., for every  $n \in V$  and every  $t \in \lambda(n)$ ,  $Out\text{-}lab\text{-}type_G^\lambda(n) \cap \delta(t) \neq \emptyset$ .

For instance,  $\lambda_2$  is a valid multi-type typing of  $G_2$  w.r.t.  $S_1$ . A *semi-valid* m-typing is a m-typing that satisfies the second condition but might violate the first one. By  $L_m(S)$  we denote the family of all graphs that have a valid m-typing w.r.t.  $S$ . For a class  $\mathcal{C}$  of bag languages by  $\text{ShEx}_m(\mathcal{C})$  we denote the class of graph languages definable under the multi-type semantics with shape expressions schemas using shape expressions in  $\mathcal{C}$  only.



## 4 Closure Properties

In this section we study the closure of **ShEx** under Boolean operations, and show that while **ShEx** are not closed under union and complement, they are closed under intersection.

When it comes to the union and complement, consider the following two schemas:

$$S_{\emptyset} : t_{\emptyset} \rightarrow \epsilon, \quad S_{\circ} : t_{\circ} \rightarrow t_{\circ}.$$

The first schema  $S_{\emptyset}$  defines the family of graphs with no edges, in particular it contains the graph  $G_{\emptyset} = (\{n_{\emptyset}\}, \emptyset)$  consisting of a single node and no edge. The second schema  $S_{\circ}$  defines the family of graphs consisting of disjoint cycles only, and in particular it contains the self-loop graph  $G_{\circ} = (\{n_{\circ}\}, \{(n_{\circ}, -, n_{\circ})\})$ . It can be easily proven that any **ShEx**  $S$  that includes both  $S_{\emptyset}$  and  $S_{\circ}$  (under either semantics) must also recognize the union of graphs  $G_{\emptyset} \cup G_{\circ} = (\{n_{\emptyset}, n_{\circ}\}, \{(n_{\circ}, -, n_{\circ})\})$ . Note, however, that  $G_{\emptyset} \cup G_{\circ}$  is not recognized by neither  $S_{\emptyset}$  nor  $S_{\circ}$ . Hence, there doesn't exist a **ShEx** capturing the union of  $S_{\emptyset}$  and  $S_{\circ}$ . Also, with a pumping argument we can show that there is no **ShEx** schema that captures the complement of  $S_{\circ}$ , again under either semantic.

**Proposition 4.1** ***ShEx<sub>s</sub>** and **ShEx<sub>m</sub>** are not closed under complement and union.*

**ShEx** are, however, closed under intersection (under both semantics). To show this non-trivial result, we need to introduce a number of auxiliary tools, which we illustrate on a simple example. Take the following two schemas defined with the following rules (recall that if the rule for a type  $t$  is missing, the default rule  $t \rightarrow \epsilon$  is in effect)

$$\begin{aligned} S_1 : t_0 \rightarrow a :: t_1^* \parallel b :: t_2^* \parallel b :: t_3^* \\ S_2 : t'_0 \rightarrow a :: t'_1^* \parallel b :: t'_2^* \parallel b :: t'_3^* \end{aligned}$$

and let  $S_1 = (\Sigma, \Gamma_1, \delta_1)$  and  $S_2 = (\Sigma, \Gamma_2, \delta_2)$ . Consider a node that has 1 outgoing  $a$ -edge and 3 outgoing  $b$ -edges i.e., a node whose local structure is described by the following bag over  $\Sigma$

$$w_0 = \{a, b, b, b\}.$$

Suppose this node is a candidate for types  $t_0$  and  $t'_0$ . Both schemas have a number of compatible assignments of types to elements of  $w$ , for instance

$$\begin{aligned} w_1 &= \{a :: t_1, b :: t_2, b :: t_2, b :: t_3\} \in \delta_1(t_0), \\ w_2 &= \{a :: t'_1, b :: t'_2, b :: t'_3, b :: t'_3\} \in \delta_2(t'_0). \end{aligned}$$

The schema  $S$  for the intersection of  $S_1$  and  $S_2$  uses pairs of types of  $S_1$  and  $S_2$  and rules that distribute properly the types. For instance, the rule for type  $(t_0, t'_0)$  needs to contain the following bag of  $\Sigma \times (\Gamma_1 \times \Gamma_2)$

$$w_J = \{a :: (t_1, t'_1), b :: (t_2, t'_2), b :: (t_2, t'_3), b :: (t_3, t'_3)\}.$$

Note that  $w_J$  has no repetitions while both  $w_1$  and  $w_2$  do. If we remove from  $w_J$  the second component over every type, we get  $w_1$ , and analogously, if we remove from  $w_J$  the first component of every type, we get  $w_2$ . This is an aggregating projection operation. On the other hand the bag  $w_J$  is a joint distribution of the bag  $w_1$  with the bag  $w_2$ . Naturally, the

shape expression for  $(t_0, t_1)$  must recognize all joint distributions of bags in  $\delta_1(t_0)$  with bags in  $\delta_2(t'_0)$ . We next formally define these two notions.

Given a bag  $w$  over  $\Sigma \times \Gamma_1 \times \Gamma_2$ , the *aggregating projection* of  $w$  on  $\Sigma \times \Gamma_1$  is a bag  $\pi_{\Sigma \times \Gamma_1}(w)$  over  $\Sigma \times \Gamma_1$

$$[\pi_{\Sigma \times \Gamma_1}(w)](a_1 :: t_1) = \sum \{w(a :: (t_1, t_2)) \mid t_2 \in \Gamma_2\}.$$

We define analogously  $\pi_{\Sigma \times \Gamma_2}(w)$  and  $\pi_{\Sigma}(w)$ . As an example,  $w_1 = \pi_{\Sigma \times \Gamma_1}(w_J)$ ,  $w_2 = \pi_{\Sigma \times \Gamma_2}(w_J)$ , and  $\pi_{\Sigma}(w_J) = w_0$ . We extend these operators to bag languages in the canonical fashion e.g.,  $\pi_{\Sigma \times \Gamma_1}(L) = \{\pi_{\Sigma \times \Gamma_1}(w) \mid w \in L\}$ .

Now, let  $w_1$  be a bag over  $\Sigma \times \Gamma_1$ ,  $w_2$  be a bag over  $\Sigma \times \Gamma_2$ ,  $L_1$  be a bag language over  $\Sigma \times \Gamma_1$ , and  $L_2$  be a bag language over  $\Sigma \times \Gamma_2$ . A bag  $w$  over  $\Sigma \times \Gamma_1 \times \Gamma_2$  is a *joint distribution* of bags  $w_1$  and  $w_2$  iff  $w_1 = \pi_{\Sigma, \Gamma_1}(w)$  and  $w_2 = \pi_{\Sigma, \Gamma_2}(w)$ . The *distributing join* of languages  $L_1$  and  $L_2$  is the set of all joint distributions of their elements:

$$L_1 \bowtie L_2 = \{w \mid \pi_{\Sigma \times \Gamma_1}(w) \in L_1 \wedge \pi_{\Sigma \times \Gamma_2}(w) \in L_2\}.$$

We point out that  $w_1$  and  $w_2$  have joint distributions if and only if  $\pi_{\Sigma}(w_1) = \pi_{\Sigma}(w_2)$ . Consequently,  $\pi_{\Sigma}(L_1 \bowtie L_2) = \pi_{\Sigma}(L_1) \cap \pi_{\Sigma}(L_2)$ .

The schema  $S$  of the intersection of  $S_1$  and  $S_2$  uses the following rule that captures the distributing join of  $\delta_1(t_0)$  and  $\delta_2(t'_0)$ :

$$(t_0, t'_0) \rightarrow (a :: (t_1, t'_1))^* \parallel b :: (t_2, t'_2))^* \parallel b :: (t_2, t'_3))^* \parallel b :: (t_3, t'_2))^* \parallel b :: (t_3, t'_3))^*.$$

Because RBEs are equivalent to Presburger formulas and the concept of distributing join can be easily defined in first-order logic with addition, we immediately obtain the following.

**Proposition 4.2** *RBE is closed under distributing join.*

The above result is instrumental in showing that shape expression schemas under both semantics are closed under intersection.

**Theorem 4.3** *ShEx<sub>s</sub> and ShEx<sub>m</sub> are closed under intersection.*

PROOF We take two schemas  $S_1 = (\Gamma_1, \delta_1)$  and  $S_2 = (\Gamma_2, \delta_2)$ , and assume that  $\Gamma_1 \cap \Gamma_2 = \emptyset$ . We define  $S_1 \cap S_2 = (\Gamma_1 \times \Gamma_2, \delta_{\cap})$ , where  $\delta_{\cap}(t_1, t_2) = \delta_1(t_1) \bowtie \delta_2(t_2)$ .

We prove that  $L_s(S_1) \cap L_s(S_2) = L_s(S_1 \cap S_2)$  and  $L_m(S_1) \cap L_m(S_2) = L_m(S_1 \cap S_2)$ . First, we take any graph  $G$  that has valid s-typings  $\lambda_1$  and  $\lambda_2$  w.r.t.  $S_1$  and  $S_2$  respectively and define the following s-typing of  $G$  w.r.t.  $S_1 \cap S_2$ :  $\lambda(n) = (\lambda_1(n), \lambda_2(n))$  for  $n \in V$ . To show that it is a valid s-typing, we note that for any  $n \in V$ ,  $\pi_{\Sigma}(\text{out-lab-type}_G^{\lambda_1}(n)) = \pi_{\Sigma}(\text{out-lab-type}_G^{\lambda_2}(n))$ , and therefore,  $\text{out-lab-type}_G^{\lambda_1}(n)$  and  $\text{out-lab-type}_G^{\lambda_2}(n)$  have joint distributions. Furthermore, all of them are contained in  $\delta_1(t_1) \bowtie \delta_2(t_2)$ , where  $\lambda(n) = (t_1, t_2)$ . It suffices to point out that by definition  $\text{out-lab-type}_G^{\lambda}(n)$  is also a joint distribution of  $\text{out-lab-type}_G^{\lambda_1}(n)$  and  $\text{out-lab-type}_G^{\lambda_2}(n)$ . Consequently,  $\text{out-lab-type}_G^{\lambda}(n) \in \delta(t_1, t_2)$ .

Next, we take any graph  $G$  that has a valid s-typing  $\lambda$  w.r.t.  $S_1 \cap S_2$  and construct s-typings  $\lambda_1$  and  $\lambda_2$  w.r.t.  $S_1$  and  $S_2$  respectively: for any  $n \in V$ , if  $\lambda(n) = (t_1, t_2)$ , then  $\lambda_1(n) = t_1$  and  $\lambda_2(n) = t_2$ . To show that  $\lambda_1$  and  $\lambda_2$  are valid s-typings, take any  $n \in V$  and let  $w = \text{out-lab-type}_G^{\lambda}(n)$ ,  $w_1 = \text{out-lab-type}_G^{\lambda_1}(n)$ , and  $w_2 = \text{out-lab-type}_G^{\lambda_2}(n)$ . Note that  $w_1 = \pi_{\Sigma, \Gamma_1}(w)$  and  $w_2 = \pi_{\Sigma, \Gamma_2}(w)$  i.e.,  $w$  is a joint distribution of  $w_1$  and  $w_2$ . By definition of distributing join of languages, we get that  $w_1 \in \delta(\lambda_1(n))$  and  $w_2 \in \delta(\lambda_2(n))$ .  $\square$

## 5 Expressive power

This section is devoted to the study of expressive power of **ShEx** in a large sense. We characterize the expressive power of **ShEx** by comparing it to standard formalisms such as first-order (FO) and monadic second-order (MSO) logics, automata on infinite trees, automata on graphs, context free graph grammars, and other means of defining graph languages. We also compare the expressive power of single-type and multi-type semantics.

We consider the first-order logic on graphs ( $\text{FO}_G$ ) over the standard signature consisting of relation names  $(E_a)_{a \in \Sigma}$ , and the existential monadic second-order logic on graphs ( $\exists \text{MSO}_G$ ) allowing only formulas of the form  $\exists X_1, \dots, X_n \varphi$ , where  $X_1, \dots, X_n$  are monadic second order variables and  $\varphi$  is a first-order logic formula using additional atomic formulae of the form  $x \in X_i$ . We say that a class of graph languages  $\mathcal{C}$  separates a graph  $H$  from a graph  $G$  if there is  $L \in \mathcal{C}$  such that  $G \in L$  and  $H \notin L$ .

Now, consider the fork  $G_\in$  and diamond  $G_\diamond$  graphs in Figure 4. We observe that single-

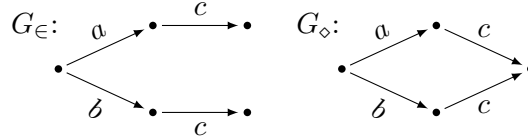


Figure 4: Fork and diamond.

type semantic can easily separate  $G_\diamond$  from  $G_\in$  while it can be easily shown that the multi-type semantic cannot. However, even the single-type semantic cannot separate  $G_\in$  from  $G_\diamond$ , while this separation can be easily accomplished with  $\text{FO}_G$  and  $\exists \text{MSO}_G$ .

Also, let  $L_{\text{cycle}}$  be the family of graphs labeled with  $\Sigma = \{a, b\}$  such that for every node  $n$  with an incoming  $b$ -edge, there is a cycle reachable from  $n$ .<sup>3</sup> It is a classic result that  $\text{FO}_G$  cannot capture families of graphs with cycles. The family  $L_{\text{cycle}}$  can be defined in both semantics with the following schema:

$$\begin{aligned} S_{\text{cycle}} : t_0 &\rightarrow (a :: t_{\text{cycle}} \mid a :: t_0)^* \parallel (b :: t_{\text{cycle}})^* \\ t_{\text{cycle}} &\rightarrow (a :: t_{\text{cycle}}^+ \mid b :: t_{\text{cycle}}) \parallel (a :: t_0)^* \parallel (b :: t_{\text{cycle}})^* \end{aligned}$$

**Proposition 5.1** *For the schema*

$$\begin{aligned} S_{\text{cycle}} : t_0 &\rightarrow (a :: t_{\text{cycle}} \mid a :: t_0)^* \parallel (b :: t_{\text{cycle}})^* \\ t_{\text{cycle}} &\rightarrow (a :: t_{\text{cycle}}^+ \mid b :: t_{\text{cycle}}) \parallel (a :: t_0)^* \parallel (b :: t_{\text{cycle}})^* \end{aligned}$$

*both  $L_s(S_{\text{cycle}})$  and  $L_m(S_{\text{cycle}})$  is exactly  $L_{\text{cycle}}$ , the family of graphs labeled with  $\Sigma = \{a, b\}$  such that for every node  $n$  with an incoming  $b$ -edge, there is a cycle reachable from  $n$ .*

**PROOF** We prove two claims that

1. If  $G$  belongs to  $L_{\text{cycle}}$ , then  $G$  has a valid s-typing w.r.t.  $S_{\text{cycle}}$ ;

Let  $V_0$  be the set of nodes of  $G$  from which no cycle is reachable. By definition of  $L_{\text{cycle}}$ , no  $b$ -labelled edge is reachable from nodes in  $V_0$ . Then the s-typing  $\lambda$  s.t.  $\lambda(n) = t_0$  if  $n \in V_0$ , and  $\lambda(n) = t_{\text{cycle}}$  otherwise is a valid s-typing for  $G$ .

<sup>3</sup>Recall that a cycle in a graph is a sequence of edges  $(n_0, a_0, n_1), (n_1, a_1, n_2), \dots, (n_k, a_k, n_0)$

2. If  $G$  has a valid m-typing w.r.t.  $S$ , then  $G$  belongs to  $L_{cycle}$ .

Let  $V_{cycle} = \{n \in V \mid t_{cycle} \in \lambda(n)\}$ . Naturally, any node with an incoming  $b$ -edge belongs to  $V_{cycle}$ . If  $V_{cycle}$  is empty, then  $G$  contains only  $a$ -labelled edges and thus belongs to  $L_{cycle}$ . More importantly, any node in  $V_{cycle}$  has at least one successor in  $V_{cycle}$ , and because  $V_{cycle}$  is finite, from any node in  $V_{cycle}$  a cycle can be reached. This proves that  $G$  belongs to  $L_{cycle}$ .  $\square$

Table 1 presents a comparison of expressive power, indicating whether a language  $L$  satisfying the constraints given in the first column can be expressed by each of the formalisms. It should

	$\text{FO}_G$	$\text{ShEx}_m$	$\text{ShEx}_s$	$\exists\text{MSO}_G$
$L : G_\diamond \in L, G_\in \in L$	✓	✓	✓	✓
$L : G_\diamond \notin L, G_\in \in L$	✓	✗	✓	✓
$L : G_\diamond \in L, G_\in \notin L$	✓	✗	✗	✓
$L_{cycle}$	✗	✓	✓	✓

Table 1: Comparison of expressive power of  $\text{ShEx}$  and the standard yardstick graph logics.

also be noted that RBEs can express cardinality constraints e.g.,  $(a\|b)^*$  means that the number of  $a$  must be equal to the number of  $b$ , that cannot be captured by  $\exists\text{MSO}_G$ .

**Corollary 5.2** *No two families of graph languages  $\text{FO}_G$ ,  $\exists\text{MSO}_G$ ,  $\text{ShEx}_s$ , and  $\text{ShEx}_m$  coincide.*

However, if we limit the expressive power of the bag languages used in schemas to those that can be captured by  $\exists\text{MSO}_G$  i.e., to the class of bag languages definable with existential monadic second-order logic on bags  $\exists\text{MSO}_B$ , the expressive power of schemas is captured by  $\exists\text{MSO}_G$ , a result easily proven with a simple adaptation of the standard translation of an automaton to an existential monadic second-order formula [?].

**Proposition 5.3**  *$\exists\text{MSO}_G$  properly contains  $\text{ShEx}_s(\exists\text{MSO}_B)$  and  $\text{ShEx}_m(\exists\text{MSO}_B)$ .*

The single-type semantic is in fact very powerful and can easily capture graph languages defined by homomorphism into a fixed graph.

**Proposition 5.4** *For any graph  $H$  there exists a  $\text{ShEx}$   $S_H$  of size  $O(|H|)$  such that*

$$L_s(S_H) = \{G \mid G \text{ is homomorphic to } H\}.$$

PROOF Let  $H = (V_H, E_H)$  and construct  $S_H = (\Sigma, V_H, \delta_H)$ , where

$$\delta_H(n) = \|(n, a, m) \in E_H (a :: m)^*$$

Now, take any  $G = (V_G, E_G)$  homomorphic to  $H$  i.e., there is a function  $h : V_G \rightarrow V_H$  s.t.  $(n, a, m) \in E_G$  implies  $(h(n), a, h(m)) \in E_H$ . It is easy to show that  $h$  is an s-typing of  $S_H$  on  $G$ . On the other hand, take a graph  $G = (V_G, E_G)$  that has a s-typing by  $S_H$ , i.e. a mapping  $\lambda : V_G \rightarrow V_H$  that satisfies the constraints of  $S_H$ . Then we can show that  $\lambda$  is a homomorphism from  $G$  to  $H$ . Indeed, let  $(n, a, m) \in E_G$ , then there exists a bag  $w \in \delta_H(\lambda(n))$  s.t.  $(a :: \lambda(m)) \in w$ . By definition of  $\delta_H$ , we deduce that  $(a :: \lambda(m))^*$  is one of the elements of the unordered concatenation in  $\delta_H(\lambda(n))$ , and therefore  $(\lambda(n), a, \lambda(m))$  is an edge in  $H$ .  $\square$

**Example 5.5** Consider the following schema which, under the single-type semantic, defines the family of graphs with homomorphism into  $K_3$  i.e., all 3-colorable graphs:

$$t_r \rightarrow - :: t_b^* \parallel - :: t_g^* \quad t_g \rightarrow - :: t_r^* \parallel - :: t_b^* \quad t_b \rightarrow - :: t_g^* \parallel - :: t_r^*$$

The multi-type semantic does not seem to be as powerful and in fact, later on, we show that the single-type semantic is strictly more expressive than the multi-type semantic.

We next briefly compare ShEx with automata on trees and graphs. A *path* in a graph is a sequence of edges  $(n_0, a_0, n_1) \cdot (n_1, a_1, n_2) \cdot \dots \cdot (n_k, a_k, n_{k+1})$ . For a node  $n$ , we denote  $Paths_n(G)$  the set of paths starting at  $n$ . A *rooted graph*  $(V, E, root)$  is a graph with a distinguished root node, and s.t. for all node  $n \in V$ , there exists a path from  $root$  to  $n$ . A *tree* is a rooted graph s.t. all nodes have in-degree one, except for the root that has in-degree of zero. The *unravelling* of a rooted graph  $G$ , denoted  $unrav(G)$ , is a possibly infinite edge-labelled tree which set of nodes is  $Paths_{root}(G)$ , which set of edges is the least set containing  $\{p, a, p \cdot (n, a, n')\}$  for all path  $p \cdot (n, a, n')$  in  $Paths_{root}G$ , and which root is the empty path  $\epsilon$ . Automata have been defined for finite or infinite ranked trees  $[?, ?]$ , and for finite unranked unordered trees  $[?]$ . Remark now that each ShEx  $S = (\Sigma, \Gamma, \delta)$  can be seen as an automaton on edge-labelled infinite trees, where  $\Gamma$  is the set of states, and  $\delta$  is the transition relation. A run is a total mapping  $\rho : Paths_{root} \rightarrow \Gamma$ , and a run is accepting if  $\delta$  is satisfied at each node. We believe that such automata would correspond to Presburger automata  $[?]$  if the latter were extended to infinite trees, taking a universal acceptance condition.

Remark now that if  $\rho$  is a valid run of  $S$  on  $unrav(G)$ , then  $\lambda$  defined by  $\lambda(n) = \{\rho(p \cdot (n', a, n)) \mid p \cdot (n', a, n) \in Paths_G(root)\}$  and  $\lambda(root) = \rho(\epsilon)$  is a valid m-tying of  $G$ . Therefore, if two graphs  $G$  and  $G'$  have isomorphic unravellings, then  $G$  and  $G'$  cannot be distinguished by any ShEx<sub>m</sub> language. This is a generalization of the unseparability by ShEx<sub>m</sub> of  $G_\infty$  and  $G_\diamond$ . Things are different for single-type semantics, as a run of  $S$  on  $unrav(G)$  cannot be transformed into a valid s-tying of  $G$ . However, if  $G$  is a finite tree, then  $unrav(G)$  is isomorphic to  $G$  and a run  $\rho$  directly yields a valid s-tying. Therefore,

**Proposition 5.6** *ShEx<sub>s</sub> and ShEx<sub>m</sub> coincide on trees: for any ShEx  $S$  and any tree  $T$ ,  $T \in L_s(G)$  if and only if  $T \in L_m(S)$ .*

Note that in the analogy between tree automata and Shape Expression Schemas, the single-type semantic is close to deterministic automata while the multi-type semantic corresponds rather to nondeterministic automata. Inspired by the classic power-set technique used to determinize a nondeterministic automaton, we extend the technique of distributing join from pairs of states to finite sets of states, and use it to show that for any given schema  $S$  interpreted under multi-type semantic one can construct a schema  $S^P$  that under single-type semantic defines the same family of graphs i.e.,  $L_m(S) = L_s(S^P)$ , albeit at the cost of increasing the number of types exponentially. Since that the multi-type semantics cannot separate  $G_\diamond$  from  $G_\infty$ , but the single-type semantics can, this allows us to state the following interesting result.

**Theorem 5.7** *ShEx<sub>s</sub> properly contains ShEx<sub>m</sub>.*

**PROOF** It suffices to show that ShEx<sub>s</sub> contains ShEx<sub>m</sub> since the proper containment follows from Table 1. First, we generalize the operations of aggregating projection and distributing joins. Given a collection of finite sets of symbols  $A_0, A_1, A_2, \dots, A_k$ ,  $w$  a bag over  $A_0 \times A_1 \times$

$\dots \times A_k$ , and a nonempty subset  $P = \{i_1, \dots, i_m\} \subseteq \{0, 1, \dots, k\}$ , the *aggregating projection* of  $w$  on  $P$  is a word  $\pi_P(w)$  over  $A_{i_1} \times \dots \times A_{i_m}$  defined as

$$[\pi_P(w)](a_{i_1}, \dots, a_{i_m}) = \sum \{w(a_1, \dots, a_k) \mid a_j \in A_j \text{ for } j \notin P\}.$$

Now, for  $i \in \{1, \dots, k\}$  let  $w_i$  be a bag over  $A_0 \times A_i$  and  $L_i$  a bag language over  $A_0 \times A_i$ . A bag  $w$  over  $A_0 \times A_1 \times \dots \times A_k$  is a *joint distribution* of bags  $w_1, \dots, w_k$  iff  $w_i = \pi_{\{0,i\}}(w)$  for  $i \in \{1, \dots, k\}$ . Note that the bags  $w_1, \dots, w_k$  have at least one joint distribution if and only if  $\pi_{\{0\}}(w_i) = \pi_{\{0\}}(w_j)$  for any two  $i, j \in \{1, \dots, k\}$ . The *distributing join* of languages  $L_1, \dots, L_k$  is the set of all joint distributions of their elements:

$$L_1 \bowtie \dots \bowtie L_k = \{w \in U_{A_0 \times A_1 \times \dots \times A_k} \mid \pi_{\{0,i\}}(w) \in L_i \text{ for } 1 \leq i \leq k\}.$$

Now, we take any schema  $S = (\Sigma, \Gamma, \delta)$  and assume some arbitrary total ordering of  $\Gamma$  that we implicitly use when enumerating subsets of  $\Gamma$ . We use the distributing join operator for a powerset construction of a shape expression  $S^P = (2^\Gamma \setminus \{\emptyset\}, \delta^P)$ . We point out, however, that a result of the distributive join on  $k$  languages over  $\Sigma \times \Gamma$  is a bag language over  $\Sigma \times (\Gamma^k)$  and we need to convert such languages to languages over  $\Sigma \times 2^\Gamma$ .

For than, take a nonempty set of types  $T \subseteq \Gamma$  and let  $E_T$  be a BRE that defines the language  $\bowtie_{t \in T} \delta(t)$ . Note that  $E_T$  is a bag expression over  $\Sigma \times \Gamma^{|T|}$ . Now, let  $E_T^P$  be the BRE obtained from  $E_T$  by replacing all occurrences of a symbol  $a :: (t_1, \dots, t_k)$  by the disjunction  $\bigvee_{\{t_1, \dots, t_k\} \subseteq T' \subseteq \Gamma} a :: T'$ . Now we can define  $\delta^P(T) = L(E_T^P)$  for any nonempty  $T \subseteq \Gamma$ .

We claim that  $L_m(S) = L_s(S^P)$ .

Take any graph  $G$  with a valid m-typing  $\lambda$  w.r.t.  $S$ . Take a node  $n \in V$  and suppose that  $\lambda(n) = \{t_1, \dots, t_k\}$ . Consequently, for every  $(a, m) \in \text{out-lab-node}_G(n)$ , there exists  $t_{i,a,m} \in \lambda(m)$  such that  $w_i = \{a :: t_{i,a,m} \mid (a, m) \in \text{out-lab-node}_G(n)\} \in \delta(t_i)$ . Note that the bag

$$w = \{a :: (t_{1,a,m}, \dots, t_{k,a,m}) \mid (a, m) \in \text{out-lab-node}_G(n)\}$$

is a join distribution of  $w_1, \dots, w_k$  and therefore belongs to  $\bowtie_i \delta(t_i)$ . Also, note that for every  $(a, m) \in \text{out-lab-node}_G(n)$ ,  $\{t_{1,a,m}, \dots, t_{k,a,m}\} \subseteq \lambda(m)$ , and by the construction of  $\delta^P$ , the bag

$$w^P = \{a :: \lambda(m) \mid (a, m) \in \text{out-lab-node}_G(n)\}$$

is in  $\delta^P(n)$ . This proves that  $\lambda$  is a valid s-typing of  $G$  w.r.t.  $S^P$ .

Now, take any graph  $G$  with a valid s-typing  $\lambda$  w.r.t.  $S^P$ . Fix a node  $n \in V$  and let  $T = \lambda(n) = \{t_1, \dots, t_k\}$ . For any  $(a, m) \in \text{out-lab-node}_G(n)$ , let  $T_{a,m} = \lambda(m)$  and note that the bag  $\{a :: T_{a,m} \mid (a, m) \in \text{out-lab-node}_G(n)\} \in \delta^P(T)$ , and therefore, for every  $(a, m) \in \text{out-lab-node}_G(n)$ , there is a sequence  $(t_{1,a,m}, \dots, t_{k,a,m})$  of elements from  $T_{a,m}$  such that the bag

$$w = \{a :: (t_{1,a,m}, \dots, t_{k,a,m}) \mid (a, m) \in \text{out-lab-node}_G(n)\}$$

belongs to  $\bowtie_i \delta(t_i)$ . Now, for every  $i \in \{1, \dots, k\}$ , the bag  $w_i = \{a :: t_{i,a,m} \mid (a, m) \in \text{out-lab-node}_G(n)\}$  is the aggregating projection  $\pi_{\{0,i\}}(w)$ , and therefore, belongs to  $\delta(t_i)$ . This shows that  $\delta$  is a valid m-typing of  $G$  w.r.t.  $S$ .  $\square$

Several notions of automata on graphs have been defined. Generally speaking, none of these automata models allows to capture graph languages definable by **ShEx**, because the latter allow to define arbitrary Presburger constraints on the outgoing edges of a node.

Therefore, only comparisons for particular restrictions of **ShEx** can be considered. Moreover, several of these automata models were studied for particular families of graphs. For instance, graph acceptors [?] are devices that associate states to graph nodes according to local logical constraints, but operate on graphs of bounded degree. E-automata [?] extend graph acceptors on arbitrary graphs and allow to capture  $\text{ShEx}(\text{RBE}(a^M, \parallel))$ . Recognizable sets of graphs [?], which expressiveness goes beyond MSO, capture  $\text{ShEx}(\exists\text{MSO}_B)$ . More recently, k-Pebble automata on graphs have been defined in [?]. Such automata are able to express FO properties, so are not comparable with **ShEx**.

Finally we compare **ShEx** with (context-free) graph grammars, see [?] for a brief overview on graph grammars and their expressive power. **ShEx** are incomparable with both node replacement (NR) grammars, and hyperedge replacement (HR) grammars. On the one hand, the language  $\{G_\diamond\}$  is definable by both HR and NR graph grammars with single initial graph and no rules. On the other hand, **ShEx** can define languages that are not definable by HR neither NR grammars. For any HR grammar  $\mathcal{G}$ , the graphs generated by  $\mathcal{G}$  are of bounded tree-width, the bound depending on  $\mathcal{G}$ . A simple **ShEx** can define the universal graph language, which clearly contains graphs of any tree-width. Also, it is known that no graph language generated by a NR grammar contains infinitely many square grids. One can easily define a **ShEx** which language contains all square grids. Note that the examples above hold for both single-type and multi-type semantics.

## 6 Validation

In this section we consider the problem of *validation*: checking whether a given graph has a valid typing w.r.t. a given **ShEx**. This problem has two parameters: 1) the kind of typing, either single-type or multi-type and 2) the class of regular bag expressions used for type definitions in the schema.

We first point out that the complexity of single-type validation for  $\text{ShEx}(\text{RBE})$  is NP-complete. The NP upper bound follows from the fact that the membership problem for RBE is in NP. For the lower bound, we recall that single-type semantics can define the language of 3-colorable graphs using very simple RBE i.e., single-type validation is NP-hard for  $\text{ShEx}(\text{RBE}(a^M, \parallel))$ .

**Theorem 6.1** *Single-type validation for  $\text{ShEx}(\text{RBE})$  is in NP-complete.*

**PROOF** It suffices to guess an s-typing  $\lambda$  and verify that it is valid. For that, it suffices to check for every node  $n \in V_G$  that  $\text{out-lab-type}_G^\lambda(n) \in \delta(\lambda(n))$ , and we recall that membership of a bag to the language of an RBE is known to be NP-complete [11].  $\square$

For the remainder of this section, we focus on multi-type validation.

### 6.1 Semi-lattice of m-typings

We begin by presenting a downward refinement method that allows to construct a unique maximal valid m-typing. Take a graph  $G = (V, E)$  and a **ShEx**  $S$ , and let  $mTyping(G, S)$  be the set of all valid m-typings of the graph  $G$  w.r.t. the schema  $S$ . We observe that  $mTyping(G, S)$  is a semi-lattice with the meet operation  $\sqcap$  and the (induced) partial order  $\sqsubseteq$

defined as follows:

$$\begin{aligned} (\lambda_1 \sqcup \lambda_2)(n) &= \lambda_1(n) \cup \lambda_2(n) \quad \text{for } n \in V, \\ \lambda_1 \sqsubseteq \lambda_2 &\text{ iff } \forall n \in V. \lambda_1(n) \subseteq \lambda_2(n). \end{aligned}$$

The refinement method goes as follows. We begin with a typing that assigns to every node the set of all types, and then we iteratively remove the types that are not satisfied. Formally, we define the one-step refinement operator on m-typings as follows (with  $n \in V$ ):

$$[Refine(\lambda)](n) = \{t \in \lambda(n) \mid Out\text{-}lab\text{-}type_G^\lambda(n) \cap \delta(t) \neq \emptyset\}.$$

Clearly, *Refine* is a monotone operator i.e.,  $Refine(\lambda) \sqsubseteq \lambda$ , and therefore, the fix-point  $Refine^*(\lambda)$  is well-defined. We claim that the procedure outlined above indeed constructs the maximal valid m-typing if one exists.

**Lemma 6.2** *For any  $\lambda \in mTyping(G, S)$ ,  $\lambda \sqsubseteq Refine^*(\lambda^\circ)$ , where  $\lambda^\circ(n) = \Gamma$  for every  $n \in V$ .*

PROOF First, we observe that if a valid (or even a semi-valid) m-typing  $\lambda$  is included in  $\lambda_0$ , the refinement does not remove the types of  $\lambda$  i.e.,

$$\forall \lambda \in mTyping(G, S). \forall \lambda_0. \lambda \sqsubseteq \lambda_0 \Rightarrow \lambda \sqsubseteq Refine(\lambda_0).$$

Since  $\lambda \sqsubseteq \lambda^\circ$ ,  $\lambda \sqsubseteq Refine^*(\lambda^\circ)$ . □

In particular,  $G$  satisfies  $S$  if and only if  $Refine^*(\lambda^\circ)$  is valid, and then,  $Refine^*(\lambda^\circ)$  is the  $\sqsubseteq$ -maximal valid m-typing of  $G$  on  $S$ . We point out that there does not necessarily exist a unique  $\sqsubseteq$ -minimal valid m-typing.

## 6.2 Complexity of Validation

Using the above refinement procedure, we show that multi-type validation is NP-complete for  $ShEx(RBE)$ , but is tractable for  $ShEx(a^M, \parallel)$ . Recall that  $RBE_1$  is a class of expressions of the form

$$(a_{1,1} \mid \dots \mid a_{1,k_1}) \parallel \dots \parallel (a_{n,1} \mid \dots \mid a_{n,k_n}).$$

Note that the test of non-emptiness of  $(*) Out\text{-}lab\text{-}type_G^\lambda(n) \cap \delta(t)$ , required by the refinement procedure, can be expressed as non-emptiness of intersection of an  $RBE_1$  expression with an RBE expression defining a type. Therefore, for a class of RBEs  $\mathcal{C}$  we identify the following decision problem:

$$INTER_1(\mathcal{C}) = \{(E_0, E) \in RBE_1 \times \mathcal{C} \mid L(E_0) \cap L(E) \neq \emptyset\}.$$

Tractability  $INTER_1$  is a necessary and sufficient condition for the tractability of multi-type validation for  $ShEx(\mathcal{C})$ . On the one hand, the refinement procedure described above reaches a fixed-point after a polynomial number of iterations. More precisely, the complexity of multi-type validity of a graph  $G = (V, E)$  w.r.t. a  $ShEx$  is in  $O(|G||S|f(G, S))$ , where  $f(G, S)$  is the (worst case) complexity of performing the test  $(*) Out\text{-}lab\text{-}type_G^\lambda(n) \cap \delta(t) \neq \emptyset$  for  $n \in V$ . On the other hand, we show that for any class  $\mathcal{C}$  of RBEs there exists a polynomial-time reduction from  $INTER_1(\mathcal{C})$  to validation for  $ShEx_m(\mathcal{C})$ .

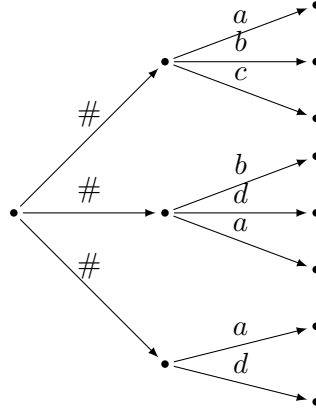


**Proposition 6.3** *For any class  $\mathcal{C}$  of BREs there exists a polynomial-time reduction from  $\text{INTER}_1(\mathcal{C})$  to validation for  $\text{ShEx}_m(\mathcal{C})$ .*

PROOF We illustrate the reduction on an example only for  $E_0 = (a \mid b \mid c) \parallel (b \mid d \mid a) \parallel (a \mid d)$  and  $E = ((a \parallel b)^* \mid c) \parallel (d \mid a)^*$ . Let  $\Sigma = \{\#, a, b, c, d\}$ ,  $\Gamma = \{t_0, t_a, t_b, t_c, t_d, t_\epsilon\}$ , and take the following schema  $S$ :

$$\begin{aligned} t_0 &\rightarrow ((\# :: t_a \parallel \# :: t_b)^* \mid \# :: t_c) \parallel (\# :: t_d \mid \# :: t_a)^* \\ t_\sigma &\rightarrow \sigma :: t_\epsilon \quad \text{for } \sigma \in \{a, b, c, d\} \\ t_\epsilon &\rightarrow \epsilon. \end{aligned}$$

The graph  $G$  is



The main claim is that  $G \in L_m(S)$  if and only if  $E_0 \cap E \neq \emptyset$ . Naturally, the reduction can be performed in polynomial time.  $\square$

This observation allows us to show that multi-type validation for  $\text{ShEx}(\text{RBE})$  is NP-complete. The hardness follows from [6] who show that testing the emptiness of intersection is NP-hard for  $\text{RBE}(a^M, \mid, \parallel)$ . For the NP upper bound, we show in the next section that  $\text{INTER}_1(\text{RBE})$  reduces to computing the integer solutions of a system of linear equations which is in NP [15].

## 7 Satisfiability of RBEs

In this section we investigate the satisfiability of RBEs extended with the intersection operator:  $L(E_1 \cap E_2) = L(E_1) \cap L(E_2)$ . More precisely, for an expression  $E$  we wish to test whether if  $L(E) \neq \emptyset$ , and then we say that  $E$  is satisfiable. Note that any expression without the intersection operator is trivially satisfied. The complexity of the analogue problem for word regular expressions with intersection is very high: PSPACE-complete [?]. We show, however, that for regular bag expressions this problem has a lesser complexity: it is NP-complete. We also identify a tractable subclass class  $\text{RBE}(a^M, \parallel, \cap)$ . Finally, we show the implications of our findings for the problem of multi-type validation.

### 7.1 The tractable subclass

We begin with a subclass  $\text{RBE}(a^M, \parallel, \cap)$  and present a simple *normal form* for expressions from this subclass.

**Proposition 7.1** *Any expression of  $\text{RBE}(a^M, \parallel, \cap)$  over  $\Delta = \{a_1, \dots, a_n\}$  is equivalent to  $a_1^{I_1} \parallel a_2^{I_2} \parallel \dots \parallel a_n^{I_n}$  for some intervals  $I_1, I_2, \dots$ , and  $I_n$ .*

For instance,  $(a \parallel a^+) \cap (a \parallel a^? \parallel a^?)$  is equivalent to  $a^{[2;\infty]} \cap a^{[1;3]}$ , which is equivalent to  $a^{[2;3]}$ . This simple example illustrate to essential equivalence rules that allow to normalize expressions:  $a^I \cap a^J = a^{I \cap J}$  and  $a^I \parallel a^J = a^{I \oplus J}$  (recall that  $[n_1, m_1] \oplus [n_2, m_2] = [n_1 + n_2, m_1 + m_2]$ ). Naturally, a normalized expression is satisfiable iff it uses only nonempty intervals. Consequently,

**Proposition 7.2** *Satisfiability of  $\text{RBE}(a^M, \parallel, \cap)$  is in PTIME.*

This results does not allow, however, to directly identify a class of regular bag expression with tractable multi-type validation because  $\text{INTER}_1$  uses expressions with the union operator. In fact, allowing an unrestricted use of the union operator leads to intractability of both the satisfiability problem [6] and the corresponding multi-type validation problem. Consequently,  $\text{INTER}_1(\text{RBE}(a^M, \parallel))$  requires a more diligent treatment, which we present next.

We reduce  $\text{INTER}_1(\text{RBE}(a^M, \parallel))$  to the circulation problem in flow networks. Recall that a flow network is a directed graph with arcs having additionally assigned the amount of flow they require (minimum flow) and the amount of flow they can accept (maximum flow). The *circulation problem* is to find a valid flow i.e., an assignment of values to arcs of the flow network so that the sum of incoming flow is equal to the sum of outgoing flow at every node. This problem has been well-studied and a number of efficient polynomial algorithm exist (cf. [9]).

We first, illustrate the construction on the example of

$$E_0 = (a \mid c) \parallel (b \mid c) \quad \text{and} \quad E = a^? \parallel b^* \parallel c.$$

The corresponding network is presented in Figure 5. The maximum flow (the minimum flow) of an arc is indicated above (below respectively). Additionally, the values of a valid flow are indicated in circles on the arcs.

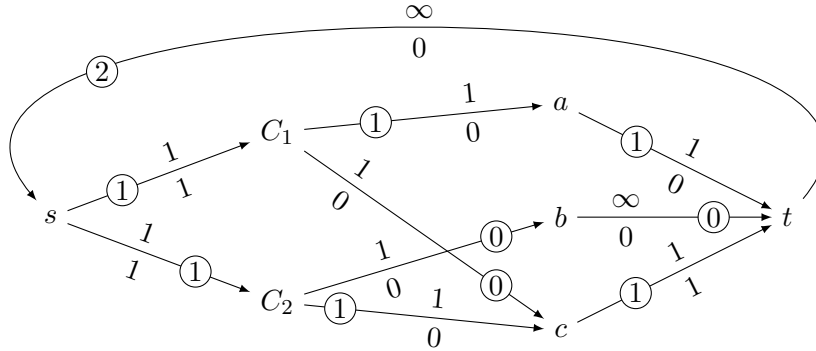


Figure 5: A flow network with a valid flow.

In general, take any  $\text{RBE}_1$   $E_0 = C_1 \parallel \dots \parallel C_k$ , where every  $C_i$  is a disjunction of symbols of  $\Delta = \{a_1, \dots, a_n\}$ , and any expression  $E \in \text{RBE}(a^M, \parallel)$ . We begin by normalizing  $E$  to  $a_1^{I_1} \parallel \dots \parallel a_n^{I_n}$ . The constructed flow  $N$  has the nodes  $V_N = \{s, c_1, \dots, c_k, a_1, \dots, a_n, t\}$  and the following arcs: 1)  $(s, c_i)$  with minimum and maximum flow 1 for every  $i \in \{1, \dots, k\}$ , 2)  $(c_i, a_j)$  with minimum flow 0 and maximum flow 1 for every  $i \in \{1, \dots, k\}$  and  $j \in \{1, \dots, n\}$  such that the disjunction  $C_i$  uses the symbol  $a_j$ , 3)  $(a_j, t)$  with minimum flow  $\min I_j$  and maximum

flow  $\max I_j$  for every  $j \in \{1, \dots, n\}$ , and 4)  $(t, s)$  with minimum flow 0 and maximum flow  $\infty$ . We claim that  $N$  has a valid flow iff  $L(E_0) \cap L(E)$  is nonempty, which shows that  $\text{INTER}_1(\text{RBE}(a^M, \parallel))$  is in PTIME, and consequently, we obtain.

**Theorem 7.3** *Multi-type validation for  $\text{ShEx}(\text{RBE}(a^M, \parallel))$  is in PTIME.*

## 7.2 The general case

For a given RBE expression with intersection, one can easily capture its satisfiability with a Presburger formula. Recall that the Presburger arithmetic is a first-order logic over natural numbers with addition  $+$ . Essentially, we recursively construct for every subexpression  $E$  a Presburger formula  $\Phi(\bar{x}, n)$  such that  $\Phi(w, n)$  if and only if  $w \in L(E^n)$ . The construction follows:

$$\begin{aligned}\Phi_\epsilon(\bar{x}, n) &:= \bigwedge_a x_a = 0, \\ \Phi_a(\bar{x}, n) &:= x_a = n \wedge \bigwedge_{b \neq a} x_b = 0, \\ \Phi_{E_1|E_2}(\bar{x}, n) &:= \exists n_1, n_2, \bar{x}_1, \bar{x}_2. n = n_1 + n_2 \wedge \bar{x} = \bar{x}_1 + \bar{x}_2 \wedge \Phi_{E_1}(\bar{x}_1, n_1) \wedge \Phi_{E_2}(\bar{x}_2, n_2), \\ \Phi_{E_1 \parallel E_2}(\bar{x}, n) &:= \exists \bar{x}_1, \bar{x}_2. \bar{x} = \bar{x}_1 + \bar{x}_2 \wedge \Phi_{E_1}(\bar{x}_1, n) \wedge \Phi_{E_2}(\bar{x}_2, n), \\ \Phi_{E_1 \cap E_2}(\bar{x}, n) &:= \Phi_{E_1}(\bar{x}, n) \wedge \Phi_{E_2}(\bar{x}, n), \\ \Phi_{E^*}(\bar{x}, n) &:= \exists n'. \Phi_E(\bar{x}, n').\end{aligned}$$

Clearly, an expression  $E$  is satisfiable if and only if  $\Psi_E = \exists \bar{x}. \Phi_E(\bar{x}, 1)$  is valid. We point out that  $\Psi_E$  is an existential-quantified conjunction of equalities of linear combinations of integer variables i.e., a system of linear equations whose only integer solutions interest us. Consequently, we can use results on integer linear programming [15] to show the following.

**Theorem 7.4** *Satisfiability of  $\text{RBE}(a^M, \parallel, |, \cap, *)$  is NP-complete.*

We point out that this result illustrates yet another computational difference between regular expressions for words and regular expressions for bags: the analogous problem for word regular expressions with intersection is known to be PSPACE-complete [?]. Naturally, this result also settles the question of the upper bound of multi-type validation for arbitrary RBEs.

**Corollary 7.5** *Multi-type validation for  $\text{ShEx}(\text{RBE})$  is NP-complete.*

## 8 Determinism

Determinism is a classical tool for decreasing the complexity of validation. In this section we are interested in the complexity of validation for deterministic ShEx. We show that multi-type validation for deterministic ShEx is not harder than membership of a bag to the language of an RBE. This allows us to identify a large and practical class of single-occurrence regular bag expressions (SORBE) that render validation tractable (Section 8.2). Then in Section 8.3 we investigate the problem of partial validation, where the conformance of only a fragment of the input graph is to be checked, which is an important practical use case. We present an optimal algorithm for partial validation which is tractable for classes of deterministic ShEx with tractable membership, for both multi-type and single-type semantics.

## 8.1 Deterministic Shape Expressions

A shape expression  $E$  is *deterministic* if every label  $a \in \Sigma$  is used with at most one type  $t \in \Gamma$  in  $E$ . For instance,  $E_1 = a::t_1 \| b::t_2^* \| a::t_1 \| c::t_2$  is deterministic but  $E_2 = a::t_1 \| b::t_2^* \| a::t_3 \| c::t_2$  is not because the symbol  $a$  is used with two different types  $t_1$  and  $t_3$ . Now, a shape expression schema  $S = (\Sigma, \Gamma, \delta)$  is *deterministic* if it uses only deterministic shape expressions, and then, by  $\delta(t, a)$  we denote the unique type used with the symbol  $a$  in the expression used to define  $\delta(t)$  (if  $a$  is used in this expression).

Recall that the tractability of the refinement method for multi-type validation presented in Section 6.1 depends on the tractability of testing that  $\text{Out-lab-type}_G^\lambda(n) \cap \delta(t)$  is nonempty for a given graph  $G = (V, E)$ , a given node  $n \in V$ , and a typing  $\lambda$  of  $G$  w.r.t. a ShEx  $S = (\Sigma, \Gamma, \delta)$  is a schema, and  $t \in \Gamma$ . When  $S$  is deterministic, the unique bag from  $\text{Out-lab-type}_G^\lambda(n)$  that can possibly belong to  $\delta(t)$  is the bag that uses the type  $\delta(t, a)$  for every label  $a$ . More precisely, if there is an edge  $(n, a, m)$  in  $G$  s.t.  $\lambda(m)$  does not contain the type  $\delta(t, a)$ , then  $\text{Out-lab-type}_G^\lambda(n) \cap \delta(t)$  is trivially empty. Otherwise, consider the word  $w = \{a :: \delta(t, a) \mid (n, a, m) \in E\}$ :  $\text{Out-lab-type}_G^\lambda(n) \cap \delta(t)$  is non empty if, and only if,  $w$  belongs to  $\delta(t)$ . More formally, we state it as follows.

**Proposition 8.1** *For a deterministic ShEx  $S$ , a (possibly invalid)  $m$ -typing  $\lambda$  of a graph  $G$  w.r.t.  $S$ , and a node  $n \in V$  we have*

$$[\text{Refine}(\lambda)](n) = \{t \in \Gamma \mid \text{out-lab}_G(n) \in \pi_\Sigma(\delta(t)) \wedge \forall (a, m) \in \text{out-lab-node}_G(n). \delta(t, a) \in \lambda(m)\}.$$

Using this argument, we establish a relationship between multi-type validation for a class  $\mathcal{C}$  of deterministic RBEs and testing the membership of a bag of symbols to the language of an RBE. Formally, for a class of RBE  $\mathcal{C}$ , by  $\text{MEMB}(\mathcal{C})$  we denote the membership problem for  $\mathcal{C}$ : given a bag of symbols  $w$  and an RBE  $E$ , check whether  $w \in L(E)$ . We point out that in this context, we can assume that the bag  $w$  represented in the unary form i.e., the size of  $w$  is  $\sum_{a \in \Sigma} w(a)$ , because it correspond to the neighborhood of a node in the graph being validated.

**Proposition 8.2** *For any  $\mathcal{C}$  of RBEs definable by deterministic shape expressions, there is a polynomial-time reduction from  $\text{MEMB}(\mathcal{C})$  to multi-type validation for  $\mathcal{C}$ .*

## 8.2 Single-occurrence RBE (SORBE)

While the problem of membership of a bag to a language defined by an RBE is in general intractable [11], we identify a rich and practical class of RBEs with tractable membership. A *single-occurrence regular bag expression* (SORBE) over  $\Delta$  is an RBE that allows at most one occurrence of every symbol of  $\Delta$  and allows the use of the wildcard  $+$  inside expressions and arbitrary intervals on symbols  $a^I$ . Note that this also enables the use of the wildcard  $?$  since it can be defined using  $\epsilon$  and the union operator without repeating any symbol of  $\Delta$ .

**Theorem 8.3**  $\text{MEMB}(\text{SORBE})$  is in  $\text{PTIME}$ .

**PROOF** We fix a bag of symbols  $w$  over  $\Delta$  and for an regular bag expression  $E$  by  $\Delta(E)$  we denote the subset of  $\Delta$  containing exactly the symbols used in  $E$ . For a subset  $X \subseteq \Delta$  by  $w_X$  we denote the bag over  $X$  obtained from  $w$  by removing all occurrences of symbols outside

of  $X$ . W.l.o.g. we assume that the Kleene's plus  $E^+$  is used only if  $\varepsilon \notin L(E)$ . Indeed, if  $\varepsilon \in L(E)$ , then  $E^+$  is equivalent to  $E^*$  and we make no restrictions on the use the Kleene's star.

The algorithm constructs for an expression an interval  $I(E)$  such that  $i \in I(E)$  iff  $w_{\Delta(E)} \in L(E)^i$ . The interval is constructed recursively as follows (with  $0/\infty = 0$  and  $i/\infty = 1$  for  $i \geq 1$ ):

$$\begin{aligned}
I(\epsilon) &= [0; \infty], \\
I(a^{[n,m]}) &= [\lceil w(a)/m \rceil; \lfloor w(a)/n \rfloor], \\
I(E_1 \mid E_2) &= I(E_1) \oplus I(E_2), \\
I(E_1 \parallel E_2) &= I(E_1) \cap I(E_2), \\
I(E^*) &= \begin{cases} [0; \infty] & \text{if } w_{\Delta(E)} = \varepsilon, \\ [1; \infty] & \text{if } w_{\Delta(E)} \neq \varepsilon \text{ and } I(E) \neq \emptyset, \\ \emptyset & \text{otherwise,} \end{cases} \\
I(E^+) &= \begin{cases} [0; 0] & \text{if } w_{\Delta(E)} = \varepsilon, \\ [1; \max I(E)] & \text{if } w_{\Delta(E)} \neq \varepsilon \text{ and } I(E) \neq \emptyset, \\ \emptyset & \text{otherwise.} \end{cases}
\end{aligned}$$

Naturally,  $w \in L(E)$  if and only if  $1 \in I(E)$  and  $w$  does not use any symbol outside of those used in  $E$ . We point out that  $I(E^+) = [0; 0]$  when  $w_{\Delta(E)} = \varepsilon$  indeed holds because of the assumption that  $\varepsilon \notin L(E)$  (otherwise the use of Kleene's star  $E^*$  in place of Kleene's plus would be enforced).  $\square$

From Proposition 8.2 and Theorem 8.3 we immediately get

**Corollary 8.4** *Multi-type validation for deterministic Shape Expressions using SORBE is in PTIME.*

We show, however, that single-type validation remains NP-complete for deterministic ShEx(SORBE).

**Theorem 8.5** *Single-type validation for deterministic Shape Expressions using SORBE is NP-complete.*

**PROOF** We present a reduction from exact set cover, a known NP-complete problem where given a set  $U = \{1, \dots, n\}$  and family  $\mathcal{S} = \{S_1, \dots, S_k\}$  of subsets of  $U$  the question is whether there exists a subfamily  $\mathcal{S}' \subseteq \mathcal{S}$  of pairwise disjoint sets that covers  $U$  i.e.,  $\bigcup \mathcal{S}' = U$ .

We illustrate the reduction on an example of  $U = \{1, 2, 3\}$  and  $\mathcal{S} = \{S_1, S_2, S_3\}$  with  $S_1 = \{1, 3\}$ ,  $S_2 = \{1, 2\}$ ,  $S_3 = \{2\}$ . We construct a schema over the set of symbols

$$\Sigma = \{1, 2, 3, S_1, S_2, S_3\}$$

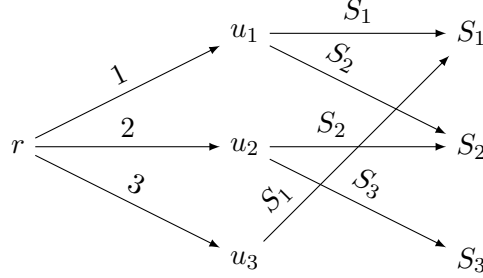
and the set of types

$$\Gamma = \{t_0, t_{1,S_1}, t_{1,S_2}, t_{2,S_2}, t_{2,S_3}, t_{3,S_1}, In, Out\}$$

that consists of the following rules.

$$\begin{aligned}
t_0 &\rightarrow (1 :: t_{1,S_1})^* \parallel (1 :: t_{1,S_2})^* \parallel (2 :: t_{2,S_2})^* \parallel (2 :: t_{2,S_3})^* \parallel (3 :: t_{3,S_1})^* \\
t_{1,S_1} &\rightarrow (S_1 :: In \parallel S_2 :: Out) \\
t_{1,S_2} &\rightarrow (S_2 :: In \parallel S_1 :: Out) \\
t_{2,S_2} &\rightarrow (S_2 :: In \parallel S_3 :: Out) \\
t_{2,S_3} &\rightarrow (S_3 :: In \parallel S_2 :: Out) \\
t_{3,S_1} &\rightarrow (S_1 :: In) \\
In &\rightarrow \epsilon \\
Out &\rightarrow \epsilon
\end{aligned}$$

Together with the following graph.



In essence  $t_{i,S_j}$  assigned to the node  $u_i$  indicates the element  $i$  is covered by set  $S_j$  and the types  $In$  and  $Out$ , assigned to the  $S_j$  nodes, are used to identify the sets that make part of the covering. For instance, the exact covering  $\{S_1, S_3\}$  corresponds to the following valid s-typing:

$$\begin{array}{lll}
\lambda(r) = t_0, & \lambda(u_1) = t_{1,S_1}, & \lambda(S_1) = In, \\
& \lambda(u_2) = t_{2,S_2}, & \lambda(S_2) = Out, \\
& \lambda(u_3) = t_{3,S_1}, & \lambda(S_3) = In.
\end{array} \quad \square$$

### 8.3 Optimal validation algorithm

Some applications might not require testing validity of the whole graph, but rather checking the validity of only a small fragment that will be accessed by the application. Such a fragment can be identified by a set of root nodes, entry points for navigating the graph, and typically, the application will require the entry points to satisfy certain types. In this section, we show how this scenario can be modeled with ShEx and present an efficient algorithm that works with deterministic shape expressions.

For this, take a schema  $S = (\Sigma, \Gamma, \delta)$  such that  $\Gamma$  contains a special *universal type* type  $t_\top$  with the definition  $\delta(t_\top) = (\Sigma \times \Gamma)^*$ . The language of  $S$  is the universal graph language, as any node of any graph can be typed with  $t_\top$ . In essence, the universal type allows to forgo validation of any node because any node implicitly satisfies  $t_\top$ .

To carry out validation on a fragment of a graph identified by the entry points we introduce the notion of pre-typing, an assignment of required types to a selected set of nodes. Formally, a *pre-typing* of a graph  $G$  (w.r.t.  $S$ ) is a partial mapping  $\lambda_- : V \rightarrow 2^\Gamma$ . Now, the objective is

to find a *valid extension* of  $\lambda_-$  i.e., a valid m-typing  $\lambda$  of  $G$  w.r.t.  $S$  such that  $\lambda_- \sqsubseteq \lambda$ . The universal type  $t_\top$  combined with pre-typing is a very powerful modelling tool. For instance, the rule  $t_0 \rightarrow a :: t_1^* \parallel (\parallel_{b \in \Sigma, b \neq a} b :: t_\top)$  indicates that we are interested in checking correct typing for nodes reachable by an  $a$ -labelled edge, but all remaining nodes can have arbitrary types, therefore do not need to be typed.

Since we are not interested in typing the whole graph  $G$ , we focus on the smallest possible valid extension of a given pre-typing  $\lambda_-$ . Interestingly, we can show the following.

**Lemma 8.6** *For a deterministic ShEx  $S = (\Sigma, \Gamma, \delta)$  with universal type, a graph  $G = (V, E)$ , and a pre-typing  $\lambda_- : V \rightarrow 2^\Gamma$ , if  $\lambda_-$  admits a valid extension, then it admits a unique  $\sqsubseteq$ -minimal valid extension.*

PROOF It suffices to show that if  $\lambda_1$  and  $\lambda_2$  are valid extensions of  $\lambda_-$ , then so is  $\lambda_1 \sqcap \lambda_2$  defined as

$$(\lambda_1 \sqcap \lambda_2)(n) = \lambda_1(n) \cap \lambda_2(n) \quad \text{for all } n \in V.$$

Indeed, let  $t \in (\lambda_1 \sqcap \lambda_2)(n)$  for some node  $n \in V$  and let  $w = \{(a, \delta(t, a)) \mid \exists n'. (a, n') \in \text{out-lab-node}_G(n)\}$ . Since  $S$  is deterministic,  $w$  is the unique possible witness that  $n$  satisfies  $t$  whatever the particular typing. Then, for  $i \in \{1, 2\}$  and for all  $(a, n') \in \text{out-lab-node}_G(n)$ , we have  $w \in \text{Out-lab-type}_G^{\lambda_i}(n)$ , which implies  $\delta(t, a) \in \lambda_i(n')$ . Thus,  $\delta(t, a) \in (\lambda_1 \sqcap \lambda_2)(n')$ , and consequently,  $w \in \text{Out-lab-type}_G^{\lambda_1 \sqcap \lambda_2}(n)$ .

We present an algorithm that constructs the minimal valid extension of a given pre-typing  $\lambda_-$  of a given graph  $G$  w.r.t. a given deterministic Shape Expression Schema  $S$  with universal type. For technical reasons, we represent a typing as binary relation between the set of nodes and the set of types, and deliberately omit the universal type. More precisely, we use a relation  $R_\lambda \subseteq V \times (\Gamma \setminus \{t_\top\})$  to represent the typing

$$\lambda(n) = \begin{cases} \{t \mid (n, t) \in R_\lambda\} & \text{if } (n, t) \in R_\lambda \text{ for some } t \in \Gamma, \\ \{t_\perp\} & \text{otherwise.} \end{cases}$$

In particular, we abuse notation and use  $\lambda$  instead of  $R_\lambda$ . Also, recall that  $\delta(t, a)$  is the unique type used together with the symbol  $a$  in the definition of  $\delta(t)$ .

This algorithm is a modified graph flooding algorithm that maintains a frontier set  $F$  of pairs  $(n, t)$  for which it remains to be verified that the node  $n$  satisfies type  $t$ . Initially this set contains only the pairs specified by the pre-typing (line 1). The algorithm fails whenever for some  $(n, t) \in F$  the outgoing edges of  $n$  do not satisfy the structural constraints given by  $\delta(t)$  (lines 5–7). If, however, the constraints are satisfied, any node  $m$  reachable from  $n$  is added to  $F$  with an appropriate type unless the type is universal, in which case we do not need to perform validation, or this node is already known to satisfy this type (lines 9–11).

Note that a run of the algorithm would consider the pair  $(t, n)$  at most once, and therefore the main loop is executed at most  $|V| \times |\Gamma|$  times. Once  $F$  is empty, the constructed  $\lambda$  represents the *minimal valid extension* of  $\lambda_-$ . This algorithm is optimal in the sense that it constructs the minimal representation of the minimal valid extension and considers assigning a type to a node only if it is required to construct the extension. Analogously to Proposition 6.3 we prove the following.

---

**Algorithm 1** Minimal Valid Extension.

---

**Input:**  $S = (\Sigma, \Gamma, \delta)$  a deterministic ShEx,

$G = (V, E)$ ,

$\lambda_- \subseteq V \times \Gamma$  a pre-typing;

**Output:**  $\lambda \subseteq V \times \Gamma$  the minimal valid extension of  $\lambda_-$ .

```
1: let  $F := \lambda_-$ 
2: let  $\lambda := \emptyset$ 
3: while  $F \neq \emptyset$  do
4:   choose  $(n, t) \in F$  and remove it from  $F$ 
5:   let  $w := \{(a, \delta(t, a)) \mid (a, m) \in \text{out-lab-node}_G(n)\}$ 
6:   if  $w \not\subseteq \delta(t)$  then
7:     fail
8:    $\lambda := \lambda \cup \{(n, t)\}$ 
9:   for  $(a, m) \in \text{out-lab-node}_G(n)$  do
10:    if  $\delta(t, a) \neq t_\top$  and  $(m, \delta(t, a)) \notin \lambda$  then
11:       $F := F \cup \{(m, \delta(t, a))\}$ 
12: return  $\lambda$ 
```

---

**Proposition 8.7** *Given a deterministic ShEx( $\mathcal{C}$ )  $S = (\Sigma, \Gamma, \delta)$ , a graph  $G = (V, E)$ , and a pre-typing  $\lambda_- : V \rightarrow \Gamma$ , the algorithm Minimal Valid Extension returns the minimal valid extension of  $\lambda_-$  if it exists, or fails otherwise. The algorithm runs in PTIME whenever MEMB( $\mathcal{C}$ ) is in PTIME.*

A slight modification of this algorithm works for the single-type semantic too: in the inner loop (lines 9–11) it suffices to add a check that neither  $F$  nor  $\lambda$  contain  $(m, t')$  for some  $t' \neq \delta(t, a)$ , which prevents assigning two different types to the same node. As a result such modified algorithm constructs an s-typing  $\lambda$  (with universal type omitted). Also, note that with the single-type modification the while loop is executed at most  $|V|$  times, and the algorithm considers each edge of the graph at most once, i.e. the algorithm makes a single pass over the graph.

## 8.4 Unamibiguity

A more general notion of determinism could be used. A bag  $w$  over  $\Sigma \times \Gamma$  is *unambiguous* iff for any two  $(a_1, t_1), (a_2, t_2) \in w$ , if  $a_1 = a_2$ , then  $t_1 = t_2$ . A shape expression  $E$  is *unambiguous* if every word in  $L(E)$  is unambiguous and there are not two  $w_1, w_2 \in L(E)$  such that  $\pi_\Sigma(w_1) = \pi_\Sigma(w_2)$ . Clearly, any deterministic shape expression is also unambiguous but the converse needs not hold. For instance,  $E_3 = (a :: t_1 \parallel b :: t_2) \mid (a :: t_3 \parallel c :: t_4)$  is unambiguous but it is not deterministic.

Interestingly, unambiguity of RBE can be defined with RBEs: if  $\Sigma = \{a_1, \dots, a_n\}$  and  $\Gamma = \{t_1, \dots, t_m\}$ , let

$$E_{\text{UNAMB}} := (a_1 :: t_1^* \mid \dots \mid a_1 :: t_m^*) \parallel \dots \parallel (a_n :: t_1^* \mid \dots \mid a_n :: t_m^*).$$

$E$  is unambiguous iff  $L(E) \subseteq L(E_{\text{UNAMB}})$ . While there exists a number of classes of RBEs that contain  $E_{\text{UNAMB}}$  and have polynomial procedures for testing the containment, in general testing whether an expression is unambiguous is intractable.



**Theorem 8.8** *Testing unambiguity of an RBE is coNP-complete.*

PROOF The proof is by reduction from the emptiness of the intersection of two RBEs, which we illustrate on an example. Let  $E_1 = (a \mid b^+) \parallel c^+$  and  $E_2 = (a \parallel b)^* \parallel c^? \parallel b$ . We construct typed expressions  $E'_1 = (a :: t_a \mid b :: t_b^+) \parallel c :: t_c^+$  and  $E'_2 = (a :: t_a \parallel b :: t_b)^* \parallel c :: t_c^? \parallel b :: t_b$  and extend them with one additional conjunct  $E''_1 = (x :: t_1) \parallel E'_1$  and  $E''_2 = (x :: t_2) \parallel E'_2$ . Clearly,  $L(E_1) \cap L(E_2) = \emptyset$  if and only if  $E''_1 \mid E''_2$  is unambiguous.

To show that the problem is in coNP we devise the following algorithm. We construct  $\Psi_E(\bar{x}) = \Phi_E(\bar{x}, 1)$  (see Theorem 7.4) and for any  $a \in \Sigma$  and  $t_1, t_2 \in \Gamma$  define the following ground formula

$$\Psi_{\text{AMB}}^{a, t_1, t_2} := \exists \bar{x}, \bar{y}. \Psi_E(\bar{x}) \wedge \Psi_E(\bar{y}) \wedge \bigwedge_{b \in \Sigma} \sum_{t \in \Gamma} x_{b:t} = \sum_{t \in \Gamma} y_{b:t} \wedge x_{a:t_1} = y_{a:t_2}.$$

As in Theorem 7.4, this formula corresponds to a system of equations with integer variables, and therefore, testing its validity is in NP. We point out that  $E$  is not unambiguous iff  $\Psi_{\text{AMB}}^{a, t_1, t_2}$  for some  $a \in \Sigma$  and  $t_1, t_2 \in \Gamma$  such that  $t_1 \neq t_2$ . A nondeterministic Turing machine guesses  $a, t_1$ , and  $t_2$  and then verifies that  $\Psi_{\text{AMB}}^{a, t_1, t_2}$  is valid.  $\square$

This result limits a possible practical uses of unambiguity since verifying that a given schema satisfies it seems in general to be rather unfeasable.

## 9 Experiments

**Setup.** All experiments have been executed on a Intel Core i7-4800MQ (4×2×2.70GHz-3.70Ghz, 4 cores, 8 threads), 8GB of RAM, and an SSD (reading speed rated at ~360MB/sec (with `hdparm -t`) with the Linux Mint 16 operating system (kernel version 3.11.0-12). Two languages have been used for implementation 1) Python 2.7.5 with `rdflib` ver. 4.1.1, and 2) Java 1.7.0\_51 (Oracle) with Apache Jena 2.11.1. The heap size of the Java Virtual Machine has been fixed to 6GB of RAM at every execution. Although the machine has a large number of cores, we intentionally *do not* take any particular advantage of it and all our algorithm implementations are programmed as *single-threaded*!

### 9.1 Graph generation

We employed a rather straightforward method to generate graphs that satisfy the given ShEx and contain exactly  $n$  URI nodes. First, we generate a set of  $n$  resource nodes and randomly associate to every node one type with uniform probability. Then, for every node, we randomly generate the outgoing edges in order to satisfy the required type of the node. The collection of outgoing edges is generated by traversing the RBE defining the type of the node and for every multiplicity choosing randomly a number in  $\{0, 1\}$  for  $?$ , in  $\{1, \dots, 15\}$  for  $+$ , and in  $\{0, 1, \dots, 15\}$  for  $*$ . These edges point either to randomly chosen nodes of appropriate types or to a literal value nodes that are generated additionally. We observe that in all generated RDF datasets the number of RDF triples is on average  $5.5n$ . The generation also identifies (in a rather expensive process) a set of access nodes that covers the whole graph. These nodes are used together with their types to construct a pretyping with a valid extension that types the complete graph. Identifying the root nodes (through reachability relation) was a very expensive and resource consuming process. In fact, we could not generate graphs with more than 2 million URI nodes. The generated graph is then exported in the standard XML format for RDF.

## 9.2 Implemented Algorithms

Initially, we have implemented a number of validation algorithms in Python and tested their performance them on relatively small RDF datasets (up to 100M URI nodes, stored in main memory). Then, we used the results of those tests to identify one algorithm to be implemented in Java in two flavors: one that stored the graph in the main memory and one that uses the Jena persistent store of Apache. The algorithms implemented in Python are:

1. *Flood* an implementation of Minimal Valid Extension algorithm (Section 8) for deterministic ShEx(SORBE).
2. *Refine* an implementation of the refinement method for deterministic ShEx(SORBE) (using the membership test, cf. Proposition 8.1)
3. *S-Refine* an optimized version of *Refine* that uses a smaller initial typing (details below).
4. *RBE<sub>0</sub>-Refine* the refinement method for arbitrary ShEx(RBE<sub>0</sub>) (Sections 6.1 and 7.1, Theorem 7.3); with applied diligence and the optimizations of *S-Refine*.

The *S-Refine* version of the *Refine* procedure begins with a typing that assigns to a node only the types whose local structure requirements are satisfied:

$$\lambda^\circ(n) = \{t \in \Gamma \mid \text{out-lab}_G(n) \in \pi_\Sigma(\delta(t))\},$$

where  $\pi_\Sigma(\delta(t))$  is the bag language over  $\Sigma$  that is defined using an expression obtained from the expression used to define  $\delta(t)$  by dropping types i.e., replacing every  $a :: t$  by  $a$ . Once the local structure requirements are verified, they do not need to be rechecked at each iteration of the refinement operation, which becomes much simpler:

$$[\text{Refine}(\lambda)](n) = \{t \in \lambda(n) \mid \forall(n, a, m) \in E. \delta(t, a) \in \lambda(m)\}.$$

As we point out later on, the *Flood* algorithm performed best and we implemented it in Java and tested two versions that differ in the way of storing (and accessing) the graph: *Java (mem.)* stores the RDF graph in main memory and *Java+Jena* stores the RDF dataset in a persistent store (SSD).

## 9.3 Test suites and protocol

We have designed two test suites. The first was to evaluate different validations algorithms we presented in this paper (implemented in Python). It uses a simplified RBE<sub>0</sub> version of the schema from Figure 2 with the type **Employee** removed and the properties **reproducedBy** and **reproducedOn** removed from the definition of the type **BugReport**. The algorithms have been tested on graphs of sizes from 20K URIs to 100K. In this suite each test has been performed only once.

The second suite of tests aims at evaluating scalability of the proposed approach and tests the two Java algorithms for graphs of sizes ranging from 100K to 1800K and following exactly the schema from Figure 2. Additionally, we compared *Java (mem.)* and *Flood* (Python) on much smaller graphs. In this suite, each test has been executed 4 times, and the measurement of the first execution discarded (it was often noticeably higher than the following ones, possibly due to system warm-up and/on jvm overhead). We report the average execution time of the

remaining three measurements. In both suites, we have measured only the time spent in the validation routine (by simply recording the current time right before and after calling the validation routine). In particular, for the memory based algorithms we do not count the time for loading the graph from the XML file into the memory and in case of *Java+Jena* the graph has been preloaded into the Jena storage prior to performing any tests.

## 9.4 Test results

The results of the first suite of tests are presented in Table 2 and Figure 6 (page 34). We observe that *Flood* is the best performing algorithm of this collection. While a straightforward implementation of the refinement method *Refine* is much slower, the simple optimizations of *S-Refine* allow to get the performance of the refinement method very close to those of *Flood*. This suggests that for deterministic RBEs, if the entry points (root nodes) cover the complete graph, using the *Flood* algorithm needs not be advantageous over *S-Refine*. The results for *RBE<sub>0</sub>-Refine* show that testing nonemptiness of intersection instead of testing the membership introduces a significant overhead. Overall, the time requirements of the presented algorithms are linear in the size of the RDF dataset.

The results of the second suite of tests are presented in Table 3 (page 35). We observe in Figure 7 (page 35) that the *Java (mem.)* implementation is an order of magnitude more efficient than the Python implementation *Flood* (in both implementations no particular effort has been put towards optimization and fine-tuning). While only graphs of size less than 1 million of nodes could be loaded into the main memory, the memory based implementation outperforms the implementation with externally stored graph by a factor of 3 – 4 (Figure 8, page 35). Validation of graph stored in external memory with almost 10 million triples is done under one minute (43 sec). Overall, Java algorithms seem to be quite efficient and *Java+Jen* looks like a promising candidate for a highly-scalable approach. Also this time the performance of the presented algorithms is linear in the size of the RDF dataset.

## 10 Extensions

In this section, we generalize the proposed graph model and the schema to allow for a number of interesting extensions, including general data values and data types, wildcards, and specification of incoming edges. The main purpose of this section is to show how the proposed framework of *ShEx* can be consistently extended to RDF graphs. While complexity considerations are not the main concern, we briefly discuss them at the end of this section.

**RDF model.** In essence, we generalize the definition of a graph by replacing  $\Sigma$ -labeled edges labeled with triples of object and allowing every objects to have a data value taken from an infinite enumerable set of data values  $\mathcal{D}$ . Formally, a *generalized  $\mathcal{D}$ -data graph* (or simply a *graph*) is a tuple  $G = (V, E, d)$ , where  $V$  is a finite set of nodes (objects),  $E \subseteq V \times V \times V$  is a set of triples, and  $d : V \rightarrow \mathcal{D}$  is a function that associates a data value to every object of  $G$ . This model is more general than RDF but we chose it for its simplicity. The values associated with elements of a triple  $(o, s, p) \in E$  can be interpreted as follows:  $d(o)$  and  $d(p)$  are data values stored in the nodes  $o$  and  $p$  while  $d(s)$  is the label of the edge that connects  $o$  and  $p$ .

Now, for a given graph  $G = (V, E, d)$  and a given node  $n \in V$  we define its *neighborhood*

as a subset of triples of  $\{\text{in}, \text{out}\} \times V \times V$  defined as follows:

$$N_G(n) = \{(\text{out}, \ell, m) \mid (n, \ell, m) \in E\} \cup \{(m, \ell, \text{in}) \mid (m, \ell, n) \in E\}.$$

**Value constraints.** We assume the existence of a language  $\mathcal{L}_{\mathcal{D}}$  that allows to identify subsets of  $\mathcal{D}$  of interest: for any expression  $e \in \mathcal{L}_{\mathcal{D}}$  the corresponding subset of  $\mathcal{D}$  is denoted by  $L(e) \subseteq \mathcal{D}$ . We do not make any assumptions on the language  $\mathcal{L}_{\mathcal{D}}$ , and in particular, it may allow to partition  $\mathcal{D}$  into subdomains of different types (integer, string, timestamp, etc.) for values we associate with objects or restrict the edge labels to those that satisfy a given regular expression. For backward compatibility, in the context of labeling edges, it is resonable to assume that for every  $c \in \mathcal{D}$ , there exists an expression  $e_c$  such that  $L(e_c) = \{c\}$ . The only operation we need to employ is testing the membership i.e., given a data value  $c \in \mathcal{D}$  and expression  $e \in \mathcal{L}_{\mathcal{D}}$  verify that  $c \in L(e)$ . In the sequel, we assume that testing the membership can be done in time polynomial in the size of the representation of  $c$  and  $e$ .

**Schema.** The generalization of the schema is two fold: 1) rather than to specify the label of the (outgoing) edges the generalized schema specifies types of objects used as label of the edges, and furthermore, both outgoing and incoming edges can be specified, and 2) the type definition additionally specifies constraints on the allowed data values associated with an object of a given type.

A *generalized ShEx schema* is a tuple  $S = (\Gamma, \delta, \gamma)$ , where  $\Gamma$  is a finite set of types,  $\delta$  is a neighbourhood type definition function that maps every type to a bag language over  $(\Gamma \times \Gamma \times \{\text{in}\}) \cup (\{\text{out}\} \times \Gamma \times \Gamma)$ , and  $\gamma$  is data value definition function that associates with every type a subset of  $\mathcal{D}$ . We can assume that  $\delta$  is defined using RBEs and  $\gamma$  is defined using the language  $\mathcal{L}_{\mathcal{D}}$ .

Given a graph  $G = (V, E, d)$  and a schema  $S = (\Gamma, \delta, \gamma)$ , a *typing* of  $G$  w.r.t.  $S$  is a binary relation  $\lambda \subseteq V \times \Gamma$ , where  $(n, t) \in \lambda$  means that  $\lambda$  associates type  $t$  to node  $n$ . Now, given a typing  $\lambda$  and a node  $n$  of  $G$  we define the following RBE<sub>1</sub> expressions:

1. the outbound typed-neighbourhood of  $n$

$$N_{\text{out}}(n, G, \lambda) = \parallel_{(n, \ell, m) \in E} \mid_{(\ell, t_\ell) \in \lambda} \mid_{(m, t_m) \in \lambda} (\text{out}, t_\ell, t_m).$$

2. the inbound typed-neighbourhood of  $n$

$$N_{\text{in}}(n, G, \lambda) = \parallel_{(m, \ell, n) \in E} \mid_{(\ell, t_\ell) \in \lambda} \mid_{(m, t_m) \in \lambda} (t_m, t_\ell, \text{in}).$$

3. the combined typed-neighbourhood of  $n$

$$N(n, G, \lambda) = N_{\text{out}}(n, G, \lambda) \parallel N_{\text{in}}(n, G, \lambda).$$

Note that  $N(n, G, \lambda)$  defines a bag language over  $(\Gamma \times \Gamma \times \{\text{in}\}) \cup (\{\text{out}\} \times \Gamma \times \Gamma)$ . Now, a typing  $\lambda$  is *valid* if the following two conditions are satisfied:

1.  $\lambda$  associates at least one type with every node i.e.,  $\forall n \in V \exists t \in \Gamma. (n, t) \in \lambda$
2. every node satisfies:
  - (a) data values type definitions i.e.,  $\forall (n, t) \in \lambda. d(n) \in \gamma(t)$

(b) neighbourhood type definitions i.e.,  $\forall(n, t) \in \lambda. L(N(n, G, \lambda)) \cap \delta(t) \neq \emptyset$ .

**Complexity of validation.** The results presented in the previous sections of this paper can be easily adapted to state the following result.

**Corollary 10.1** *Validating generalized data graphs against generalized ShEx schema using RBE expressions from a class  $\mathcal{C}$  is in NP and is in PTIME if  $\text{INTER}_1(\mathcal{C})$  is in PTIME.*

In particular validation remains in PTIME if only expressions in  $\text{RBE}(a^M, \parallel)$  are used to define  $\delta$ .

## 11 Related Work

Only few formalisms for graph validation have been proposed up to now. OSLC Resource Shapes (ResSh) [17] are very closed to ShEx, except for the semantics of validation. The W3C standards RDF Schema (RDFS) [5] and Web Ontology Language (OWL) [7] also allow to define graph vocabularies. In what follows, we compare ShEx to RDFS, ResSh and OWL.

We start by a general observation. It is standard in databases to accompany the concrete data (tuples, XML tree, RDF graph) by declarative (logical) rules that provide additional information on the data. Two typical examples are constraints and ontologies. Constraints are used to certify that the collection of facts composing the data satisfies some properties, e.g. integrity constraints, functional dependencies, XML schema. Given a database instance and a constraint, the data either satisfies the constraint, or does not satisfy it. Ontologies are fundamentally different, as they are not used to describe the data in a database, but to describe real-world entities, and their relationships, that are supposed to be captured by that data. The ontology rules are typically used as inference rules. Given an *extensional* (concrete) database instance, by applying the ontology rules one infers implicit facts which, together with the concrete data, represent the *intentional* database instance. We are going to use this general observation in order to compare ShEx and ResSh with RDFS and OWL. More precisely, ShEx and ResSh are clearly used for defining constraints; OWL is an ontology language, whereas RDFS could be used for both, but is not very expressive as a constraint language. This also gives the reason why one needs to define schema formalisms for graphs, such as ShEx and ResSh, instead of using the already existing standards RDFS and OWL.

We start by an informal introduction to RDFS, using a simplified version of our introductory example from Figure 1.<sup>4</sup> Figure 9 depicts an RDF graph containing RDF Schema information (the dotted-and-grayed part). RDF Schema defines the classes `rdfs:Class` and `rdf:Property`, which intuitively identify two possible uses of IRIs, as graph nodes and as edge labels, respectively. Using the properties `rdfs:subClassOf` and `rdf:subPropertyOf`, one can specify a type hierarchy. RDF nodes can have one or several types, specified by `rdf:type` edges, whereas the type of an edge is determined by its label. On the example, there are two node types, namely `BugReport` and `User`, and two edge types, namely 'related' and 'reportedBy'. Moreover, the `rdfs:domain` and `rdfs:range` properties allow to specify types for the source and target nodes of an edge of given type. On the example, the source node of `reportedBy` edges is of type `BugReport`, whereas the target node is of type `User`. Note that

<sup>4</sup>Remind that RDF data model defines data as a set of triples, which is slightly different from the edge labelled graphs that we use here. Our introduction on RDFS is intuitive, and adapted to the model of edge labelled graphs.

all the type information is part of the graph: on the example, the dotted edges and the grayed nodes are part of the RDF graph.

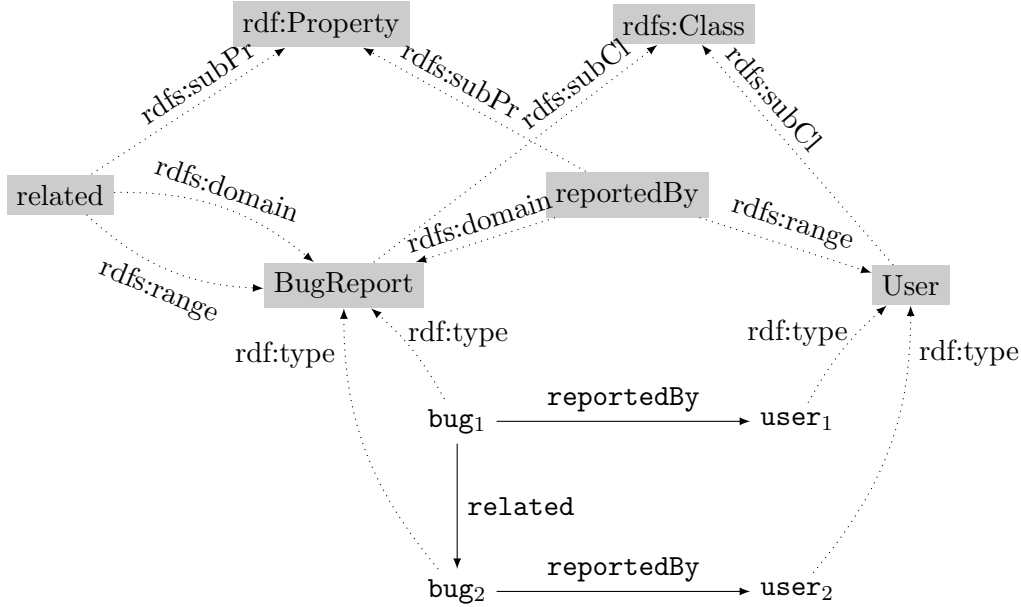


Figure 9: RDF graph with RDFS information.

The W3C RDF Schema recommendation does not fix a semantics: applications are free choose how they want to use the typing information present in a graph. The formal recommendation however suggest two possible usages of RDFS. One possibility is to use it as a light ontology, and it has been formalized by the W3C with so called entailment regimes [?]. Using that semantics on the example, if the (user1, rdfs:type, User) edge was missing in the graph, then the ontology reasoner would deduce it in order to satisfy the range requirement for the 'reportedBy' edge label. Another possibility, which is suggested by the W3C but was not formalized, is to consider RDFS type information as constraints. In that case, missing (user1, rdfs:type, User) edge should raise an error. One can imagine a third usage of RDFS, in which a type is specified for all nodes and edges in the graph, as on Figure 9. In that case, the type part of the RDF graph defines an exhaustive vocabulary for edges, and therefore restricts the admissible outgoing and incoming edges' labels for every node, however no cardinality constraints can be specified. On the example, a node of type BugReport could only have 'related' and 'reportedBy' outgoing edges. With the above-sketched semantics for RDFS, one can define a simple schema for an RDF graph. If we suppose that the RDFS type hierarchy is flat, such a schema could be captured by rules of the form  $(a_1 :: t_1^* \parallel \dots \parallel a_k :: t_k^*)$ . It is an interesting open question whether ShEx with inheritance and edge types can capture such RDFS definable schemas in presence of complex type hierarchies.

OWL is an ontology language much more expressive than RDFS-definable ontologies. It allows for instance to specify the functionality of some edge label, meaning that every node can have at most one outgoing edge for that label. This however remains an ontology, inference rule, and not a constraint. Considering the example from Figure 9, if the 'reportedBy' edge label is functional and the graph contains the two edges (bug1, reportedBy, user1) and (bug2,

reortedBy, user2), then the ontology reasoner would deduce that user1 and user2 represent the same real-words entity, which concretely is captured by an inferred intentional edge (user1, sameAs, user2). [?] proposed an alternative semantics for OWL, in which the ontology rules are considered both as inference rules and as constraints, in which the above situation would lead to an error. However, as pointed in [17], OWL with its standard semantics is widely accepted, and using it with a different semantics might be risky from the point of view of users' understanding of the system.

As we saw, RDFS and OWL ontologies are not appropriate for defining constraints, and therefore cannot play the role of a graph schema language. RDFS can be used as a constraint language, however it is unable to define cardinality constraints. The latter motivated the introduction of OSLC Resource Shapes. ResSh allow to associate cardinality constraints to RDFS types. Such constraints range over  $?, *, +, 1$  with the usual meaning, and specify the allowed number of outgoing edges of a given label. The semantics is that every node has to satisfy the constraints attached to its types. Compared to ShEx, a validation algorithm for ResSh is simpler, as it does not need to guess the possible types of a node. On the other hand, ResSh are not able to enforce constraints on graphs that do not contain RDF Schema information. Moreover, the constraints on the graph structure definable by ResSh are strictly less expressive than those definable by ShEx(SORBE). Complexity of validation for ResSh was not formally stated, but it is easily shown that it is polynomial.

In the current paper we also study complexity of membership and satisfiability for subclasses of regular bag expressions. It is related to similar studies for schemas for unordered XML [2] and XML Schemas with interleaving [6]. Our study of membership problem is related to a more general study of the membership problem of Parikh images of various families of word languages studied for instance in [10].

## 12 Conclusions

We have presented Shape Expressions Schemas (ShEx), a novel formalism of schemas for RDF graphs inspired by existing schema formalisms for semi-structured databases (XML Schema). We have proposed two alternative semantics, single-type and multi-type, studied their expressive powers and the complexity of the problem of validating a given graph w.r.t. the given schema (in the chosen semantic). In general, single-type validation is intractable while multi-type validation is tractable for a number of practical subclasses of RBEs. We have also considered a special version of the validation problem, where we only check the validity of a fragment of a graph w.r.t. a given partial pre-typing, and have shown that for the class of single-occurrence RBEs this problem can be solved efficiently for both the single- and multi-type semantics. Summary of complexity results can be found in Table 4.

	RBE( $a^M, \parallel$ )	RBE	SORBE	SORBE det.	SORBE det. + $\lambda_-$ + $t_\top$
multi-type	PTIME (Thm. 7.3)	NP-c. (Cor. 7.5)	PTIME (Cor. 8.4)		
single-type	NP-c. (Thm. 6.1)		NP-c. (Thm. 8.5)		PTIME (Section 8.3)

Table 4: Summary of main complexity results for the validation problem.

To obtain these results we have identified and studied critical problems on RBEs: the satisfiability of RBEs with intersection and the membership of a bag to a language of an

RBEs. We have shown that satisfiability of RBEs with intersection is NP-complete, a result that stands in contrast with the analogous problem for regular expressions, which is known to be PSPACE-complete. We have also identified a large class of single-occurrence RBEs with tractable membership problem.

**Future work.** We identify a number of possible directions of future work. While it is relatively easy to show with a simple reduction from containment of tree automata that the problem of testing containment of ShEx is EXPTIME-hard, its exact complexity is an open question and it would be interesting to see if determinism can be used to lower it. Naturally, it would also be interesting to explore any connections between containment of ShEx and the containment of RBEs. However, the exact complexity of containment of RBEs is an open question but it is known to be  $\Pi_2^P$ -hard and in 3EXPTIME [3]. We would also like to investigate learning algorithms for ShEx drawing inspiration from techniques based on automata learning [13, 12] as well as techniques learning regular expressions and queries [1, 18]. Another interesting topic is incremental validation of graphs w.r.t. ShEx.

## References

- [1] G. J. Bex, F. Neven, T. Schwentick, and S. Vansumneren. Inference of concise regular expressions and DTDs. *ACM Transactions on Database Systems (TODS)*, 35(2), 2010.
- [2] I. Boneva, R. Ciucanu, and S. Staworko. Simple schemas for unordered xml. In *WebDB*, pages 13–18, 2013.
- [3] I. Boneva, R. Ciucanu, and S. Staworko. Schemas for unordered XML on a DIME. Technical Report arXiv:1311.7307, arxiv.org, 2013. In preparation for journal submission. Available on at <http://arxiv.org/abs/1311.7307>.
- [4] I. Boneva, J. Emilio Labra Gayo, S. Hym, E. G. Prud’hommeau, H. Solbrig, and S. Staworko. Validating RDF with Shape Expressions. *ArXiv e-prints*, April 2014. Available at <http://arxiv.org/abs/1404.1270>.
- [5] D. Brickley and R. V. Guha. RDF Schema 1.1. <http://www.w3.org/TR/rdf-schema>, February 2004.
- [6] D. Colazzo, G. Ghelli, and C. Sartiani. Efficient inclusion for a class of XML types with interleaving and counting. *Inf. Syst.*, 34(7):643–656, 2009.
- [7] M Dean and M. Schreiber. OWL Web Ontology Language Reference. <http://www.w3.org/TR/owl-ref>, February 2004.
- [8] S. Ginsburg and Spanier E. H. Semigroups, presburger formulas, and languages. *Pacific Journal of Mathematics*, 16(2):285–296, December 1966.
- [9] A. V. Goldberg, E. Tardos, and R. E. Tarjan. Network flow algorithms. In *Algorithms and Complexity*, Volume 9, *Paths, Flows, and VLSI-Layout*, 1990.
- [10] O. H. Ibarra and B. Ravikumar. On the Parikh membership problem for FAs, PDAs, and CMs. In *Language and Automata Theory and Applications (LATA)*, pages 14–31. Springer, 2014.



- [11] E. Kopczynski and A. To. Parikh images of grammars: Complexity and applications. In *LICS*, pages 80–89, 2010.
- [12] A. Lemay, S. Maneth, and J. Niehren. A learning algorithm for top-down XML transformations. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 285–296, 2010.
- [13] J. Oncina and P. Gracia. Inferring regular languages in polynomial update time. In *Pattern Recognition and Image Analysis*, 1991.
- [14] D. C.O Oppen. A  $2^{2^{pn}}$  upper bound on the complexity of presburger arithmetic. *Journal of Computer and System Sciences*, 16(3):323–332, 1978.
- [15] C. H. Papadimitriou. On the complexity of integer programming. *Journal of the ACM*, 28(4):765–768, October 1981.
- [16] R. J. Parikh. On context-free languages. *Journal of the ACM*, 13(4):570–581, 1966.
- [17] A. Ryman, A. Le Hors, and S. Speicher. Oslc resource shape: A language for defining constraints on linked data. In *Proceedings of the WWW2013 Workshop on Linked Data on the Web (LDOW)*. CEUR-WS.org, 2013.
- [18] S. Staworko and P. Wiecek. Learning twig and path queries. In *International Conference on Database Theory (ICDT)*, March 2012.

Size		Python			
# URIs	# triples	Flood	Refine	S-Refine	RBE <sub>0</sub> -Refine
20000	110364	2.250	7.344	2.904	9.844
40000	219681	4.498	14.812	5.943	21.879
60000	328499	7.005	22.767	9.841	36.362
80000	438307	9.555	30.025	12.472	55.981
100000	548789	11.787	37.110	15.155	75.484

Table 2: Evaluation of Python implementations (times in seconds).

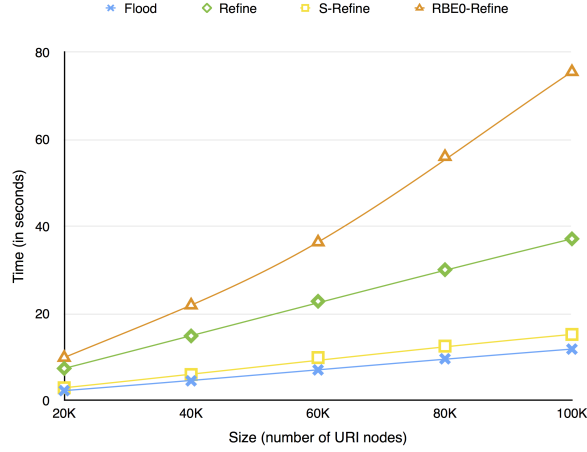


Figure 6: Evaluation times of Python implementations.

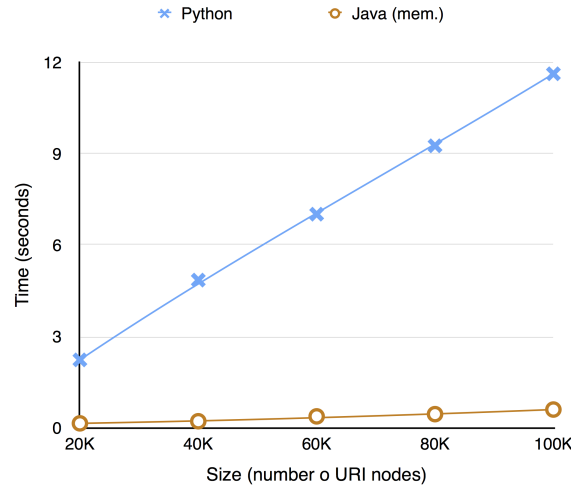


Figure 7: Comparison of Java and Python implementations (Flood).

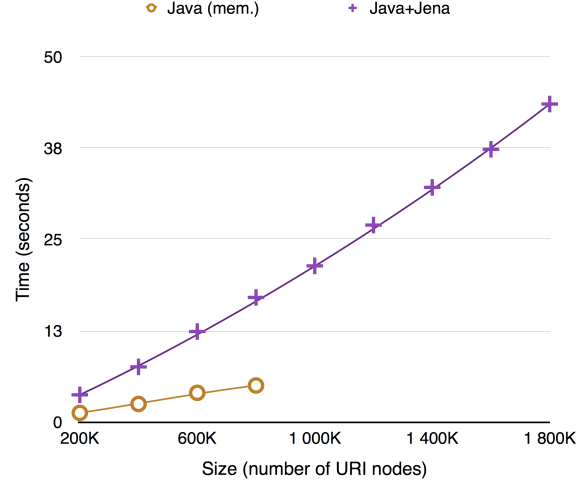


Figure 8: Comparison of memory-model flavors: Java (mem.) – main memory (RAM), Java+Jena – persistent external memory (SSD).

Size		Java		Python
# URIs	# triples	Main Memory	Jena	Flood
20000	110364	.140		2.223
40000	219681	.209		4.841
60000	328499	.369		7.004
80000	438307	.442		9.249
100000	548789	.593	1.612	11.612
200000	1095188	1.225	3.697	
400000	2202548	2.448	7.526	
600000	3292087	3.974	12.362	
800000	4394620	4.978	17.031	
1000000	5491193		21.353	
1200000	6588541		26.936	
1400000	7684722		32.102	
1600000	8778297		37.310	
1800000	9881081		43.512	

Table 3: Minimal Valid Extension (times in seconds).